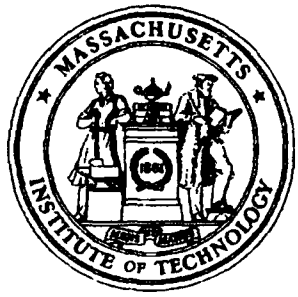
(2)



Formalizing Knowledge Used in Spectrogram Reading: Acoustic and Perceptual Evidence From Stops

RLE Technical Report No. 537

December 1988

Lori Faith Lamel

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Formalizing Knowledge Used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops

RLE Technical Report No. 537

December 1988

Lori Faith Lamel

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

This work has been supported by the Defence Advanced Research Projects Agency, Vinton-Hayes, Bell Laboratories (GRPW), and Inference Corporation.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

OSP 95294

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARPA Order No. 4585
6a. NAME OF PERFORMING ORGANIZATION: Research Laboratory of Electronics, Massachusetts Institute of Technology; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: Advanced Research Projects Agency
6c. ADDRESS (City, State, and ZIP Code): 77 Massachusetts Avenue, Cambridge, MA 02139	7b. ADDRESS (City, State, and ZIP Code): 1400 Wilson Blvd., Arlington, VA 22217
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Office of Naval Research Math. & Physical Sciences Res.; 8b. OFFICE SYMBOL (If applicable): Program	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-82-K-0727
8c. ADDRESS (City, State, and ZIP Code): 800 North Qunicy Street, Arlington, VA 22217	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. NR-049-542 / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification): Formalizing Knowledge Used in Spectrogram Reading: Acoustic and Perceptual Evidence From Stops

12. PERSONAL AUTHOR(S): Lori F. Lamel

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM _____ TO _____	December 1988	185

16. SUPPLEMENTARY NOTATION: Technical Report 537, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1988.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Please see next page

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL: Elisabeth Colford - RLE Contract Reports	22b. TELEPHONE (Include Area Code): (617)258-5871; 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE: UNCLASSIFIED

19. ABSTRACT

Since the invention of the sound spectrograph in 1946 by Koenig, Dunn and Lacey, spectrograms have been widely used for speech research. Over the last decade there has been revived interest in the application of spectrogram reading toward continuous speech recognition. Spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. While early attempts at spectrogram reading met with limited success (Klatt and Stevens, 1973; Lindblom and Svenssen, 1973; Svenssen, 1974), Zue, in a series of experiments intended to illustrate the richness of phonetic information in the speech signal (Cole et al., 1980; Cole and Zue, 1980), demonstrated that high performance phonetic labeling of a spectrogram could be obtained.

In this thesis a formal evaluation of spectrogram reading was conducted in order to obtain a better understanding of the process and to evaluate the ability of spectrogram readers. The research consisted of three main parts: an evaluation of spectrogram readers on a constrained task, a comparison to listeners on the same task, and a formalization of spectrogram-reading knowledge in a rule-based system.

The performance of 5 spectrogram readers was assessed using speech from 299 talkers. The readers identified stop consonants which were extracted from continuous speech and presented in the immediate phonemic context. The task was designed so that lexical and other higher sources of knowledge could not be used. The averaged identification rate of the ranged across contexts, from 73-82% top choice, and 77-93% for the top two choices. The performance of spectrogram readers was, on the average, 10% below that of human listeners on the same task. Listeners had an overall identification rate that ranged from 85 to 97%. The performance of readers is comparable to other spectrogram reading experiments reported in the literature, however the other studies have typically evaluated a single subject on speech spoken by a small number of talkers.

Although researchers have suggested that the process can be described in terms of rules (Zue, 1981), few compilations of rules or strategies exist (Rothenberg, 1963; Fant, 1968, Svenssen, 1974). In order to formalize the information used in spectrogram reading, a system for identifying stop consonants was developed. A knowledge-based system was chosen because the expression and use of the knowledge is explicit. The emphasis was on capturing the acoustic descriptions and modeling the reasoning thought to be used by human spectrogram readers. However, the implemention was much harder than had been anticipated due to a variety of reasons. The most important is that there appears to be much more happening in our visual system and in our thought processes than we actually express, even when asked to explain our reasoning. Human are able to selectively pay attention to acoustic evidence, even in the presence of contradictory evidence. This ability is not well understood and is difficult to mimic. The performance of the system was adequate: identification of 94 tokens that were both heard and read correctly was 88% top choice, and 96% top 2.

Formalizing Knowledge used in Spectrogram Reading:
Acoustic and perceptual evidence from stops
by
Lori Faith Lamel

Submitted to the Department of Electrical Engineering and
Computer Science on May 10, 1988 in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Abstract

Since the invention of the sound spectrograph in 1946 by Koenig, Dunn and Lacey, spectrograms have been widely used for speech research. Over the last decade there has been revived interest in the application of spectrogram reading toward continuous speech recognition. Spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. While early attempts at spectrogram reading met with limited success (Klatt and Stevens, 1973; Lindblom and Svenssen, 1973; Svenssen, 1974), Zue, in a series of experiments intended to illustrate the richness of phonetic information in the speech signal (Cole et al., 1980; Cole and Zue, 1980), demonstrated that high performance phonetic labeling of a spectrogram could be obtained.

In this thesis a formal evaluation of spectrogram reading was conducted in order to obtain a better understanding of the process and to evaluate the ability of spectrogram readers. The research consisted of three main parts: an evaluation of spectrogram readers on a constrained task, a comparison to listeners on the same task, and a formalization of spectrogram-reading knowledge in a rule-based system.

The performance of 5 spectrogram readers was assessed using speech from 299 talkers. The readers identified stop consonants which were extracted from continuous speech and presented in the immediate phonemic context. The task was designed so that lexical and other higher sources of knowledge could not be used. The averaged identification rate of the ranged across contexts, from 73-82% top choice, and 77-93% for the top two choices. The performance of spectrogram readers was, on the average, 10% below that of human listeners on the same task. Listeners had an overall identification rate that ranged from 85 to 97%. The performance of readers is comparable to other spectrogram reading experiments reported in the literature, however the other studies have typically evaluated a single subject on speech spoken by a small number of talkers.

Although researchers have suggested that the process can be described in terms of rules (Zue, 1981), few compilations of rules or strategies exist (Rothenberg, 1963; Fant, 1968, Svenssen, 1974). In order to formalize the information used in spectrogram reading, a system for identifying stop consonants was developed. A knowledge-based system was chosen because the expression and use of the knowledge is explicit. The emphasis was on capturing the acoustic descriptions and modeling the reasoning thought to be used

by human spectrogram readers. However, the implemention was much harder than had been anticipated due to a variety of reasons. The most important is that there appears to be much more happening in our visual system and in our thought processes than we actually express, even when asked to explain our reasoning. Human are able to selectively pay attention to acoustic evidence, even in the presence of contradictory evidence. This ability is not well understood, and is difficult to mimic. The performance of the system was adequate: identification of 94 tokens that were both heard and read correctly was 88% top choice, and 96% top 2.

Thesis Supervisor: Dr. Victor W. Zue
Title: Principal Research Scientist

Acknowledgments

There are many people who have helped and supported me during this thesis work and my time at MIT. I especially want to express my gratitude to:

My thesis advisor, Victor Zue, for teaching me how to read spectrograms, and providing me with the opportunity to learn from his expertise in speech and acoustic-phonetics. He has continually supported me, with guidance and friendship; Victor believed in me at times when I no longer believed in myself.

The members of my thesis committee, Ken Stevens, Stephanie Seneff, and Ramesh Patil for their interest in my work, helpful suggestions, and encouragement.

Ken Stevens and the present and past members of the Speech Communications Group for providing a stimulating environment in which to conduct research. Stephanie and I have worked closely together on many projects; I hope that we have the opportunity to do so again.

Jerry Roylance for convincing me that this was a worthwhile thesis topic and for being a friend.

Jim Glass, Caroline Huang, John Pitrelli, Stephanie Seneff, and Victor Zue, for reading spectrograms for me.

Stefanie Shattuck-Hufnagel for helping to design the perceptual experiments and for discussions and comments on an early draft.

Others who carefully reviewed drafts of this document including Nancy Daly, Susan Dubois, Carol Espy-Wilson, Pat O'Keefe and John Pitrelli, and particularly Corine Bickley, for giving me immediate feedback on my incomprehensible drafts.

Dave Whitney and Rob Kassel for making the laser writers work. I cannot thank Rob enough for providing technical assistance with the Macintosh, Latex, and the Snarfer, and for answering my continual barrage of questions, but maybe some sushi dinners and ski weekends will help!

Keith North for keeping things running smoothly and Dave Shipman, Scott Cyphers, and David Kaufman, Hong Leung, Mark Randolph and others for developing software and maintaining the lisp machines.

All of my friends who have put up with me and given my lots of encouragement, especially Corine, Michele Covell, Sue, Dan Huttenlocher, Katy Kline, Pat, Mark, Jerry, Jean-Pierre Schott and Jon Taft. Mark for long discussions late into the night, Dan for finishing his thesis which provided much of the motivation for me to get done, and Jim who commiserated with me night after night at the office while we worked on our theses.

The Zue/Seneff's for making me part of their family.

And last, but not least, I thank my family for their never ending and unquestioning love.

This work was supported DARPA, Vinton-Hayes, Bell Laboratories (GRPW). Inference Corp. provided the ART software free of charge, and Tom Goblick at Lincoln Labs generously allowed me use of their facilities before my software arrived.

Contents

List of Figures

List of Tables

Chapter 1

Spectrograms and Spectrogram Reading

While spectrograms have been used in speech analysis for many years, over the last decade there has been revived interest in the application of spectrogram reading toward continuous speech recognition. Spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. Early attempts at spectrogram reading met with limited success (Klatt and Stevens, 1973; Lindblom and Svenssen, 1973; Svenssen, 1974). In a series of experiments intended to illustrate the richness of phonetic information in the speech signal (Cole et al., 1980; Cole and Zue, 1980), Zue demonstrated that high performance phonetic labeling of a spectrogram could be obtained without the use of higher level knowledge sources such as syntax and semantics. The phonetic transcription thus obtained was better than could be achieved by automatic speech recognition phonetic front ends (Klatt, 1977). It appears that the humans' ability to handle partial specification, integrate multiple cues, and properly interpret conflicting information contributes greatly to this high level of performance.

Recently, several attempts have been made to build automatic speech recognition systems that model spectrogram reading directly (Carbonell et al., 1984; Johnson et al., 1984; Stern et al., 1986). While the attempts have met with some success, they may be somewhat premature. The spectrogram reading experiments reported in the literature have typically evaluated a single spectrogram reader on speech spoken by a small number of talkers. High performance at spectrogram reading across a large number of talkers has yet to be demonstrated. Although expert spectrogram readers have suggested that the process can be described in terms of rules (Zue, 1981), few compilations of rules or

strategies exist (Rothenberg, 1963; Fant, 1968, Svenssen, 1974). A better understanding of spectrogram reading and a more extensive evaluation is needed before computer implementations can be expected to meet with success.

In this thesis a rigorous investigation of spectrogram reading is described. The aim of the investigation was to conduct a formal evaluation of spectrogram reading in order to obtain a better understanding of the process. To do so, the performance of several spectrogram readers was assessed using speech from a large number of talkers. The task was designed so that lexical and other higher sources of knowledge could not be used. The performance of the spectrogram readers was compared to that of human listeners on the same constrained task.

Finally, an attempt was made to formalize the knowledge used in spectrogram reading by incorporating it in a knowledge-based system. The knowledge is encoded in terms of descriptions of acoustic events visible in the spectrogram, and in the relation of the acoustic events to phonemes. The relations between phonemes and acoustic events are expressed in a set of rules. Researchers have designed rule-based (or heuristic) speech recognition systems (Lesser et al., 1975; Weinstein et al., 1975; Woods et al., 1976; Erman and Lesser, 1980; Espy-Wilson, 1987); however, this formulation also attempts to model the reasoning expressed by spectrogram readers.

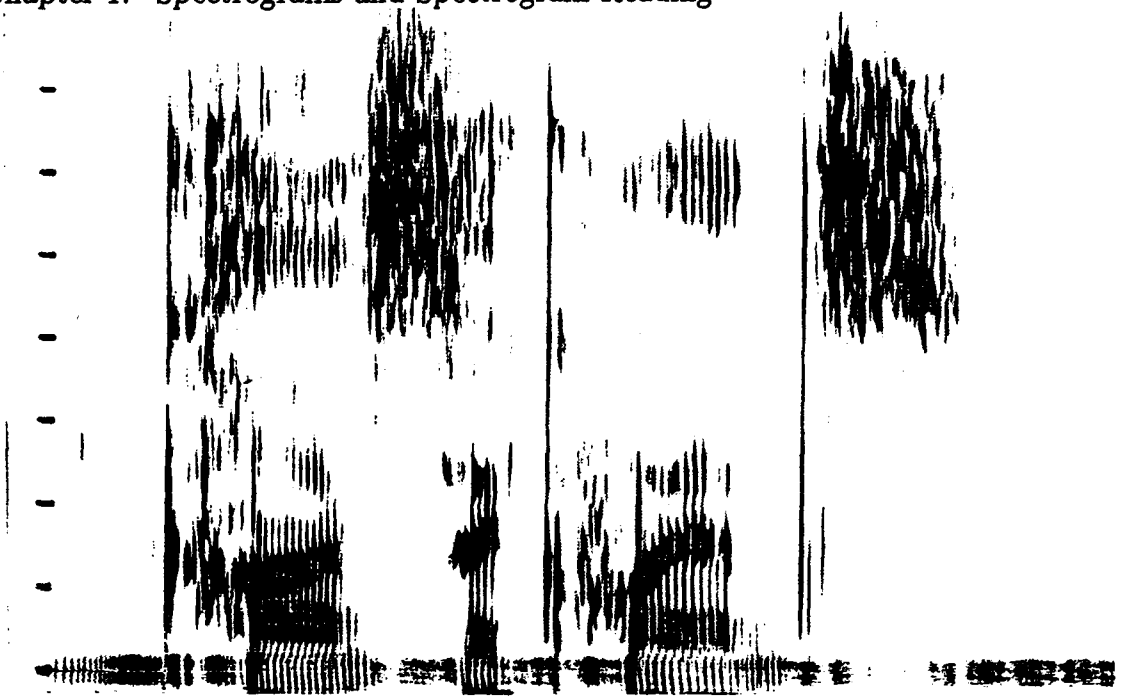
The remainder of this chapter is as follows. The first section describes spectrograms and how they are produced. Next spectrogram reading and its applications are discussed, followed by the interpretation of a spectrogram of an unknown utterance in section 1.3. Section 1.4 provides a summary of previous spectrogram reading experiments. The final section outlines the scope of this thesis.

1.1 Spectrograms

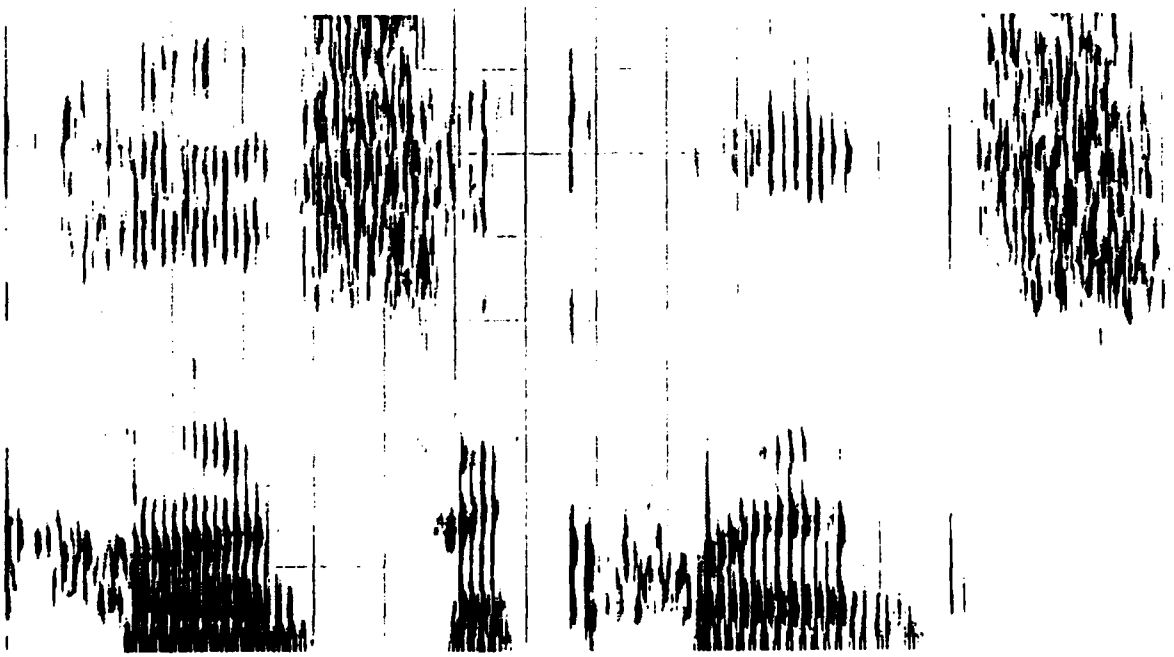
Since the invention of the sound spectrograph (Koenig, Dunn, and Lacey, 1946), spectrograms have been used extensively by researchers in the speech community. Researchers have used spectrograms to study the acoustic characteristics of speech sounds for a variety of applications, such as in the analysis of speech production and perception, in speech synthesis, to aid in automatic speech recognition and to develop aids for the handicapped. The spectrogram displays the energy distribution in the speech signal as a function of

both time and frequency. In the original implementation, an analog filter-bank was used to perform the analysis. The average energy at the output of the filters is an approximation to the short-time Fourier transform (see Equation 1.1). Koenig et al. demonstrated the effects of varying the bandwidth of the analysis filter. Two bandwidths, 45 Hz and 300 Hz, have remained the most popular. The narrow-band spectrogram, produced with a filter bandwidth of 45 Hz, is able to resolve the individual harmonics in the spectrum, and has been used primarily to measure fundamental frequency. The wide-band spectrogram, produced with a 300 Hz bandwidth, provides a convenient visual display of the acoustic characteristics of speech sounds. Since the wide-band spectrogram is produced with a short time window, it provides good temporal resolution, enabling accurate location of events in time (such as stop releases or the onset of voicing). In addition, formant frequencies and the spectral energy in noise-like regions are generally easy to resolve. The wide-band spectrogram has been used in this research.

While spectrograms are a convenient representation, some aspects of speech known to be important, such as stress and intonation, are not well represented. In addition, the analysis makes no attempt to model the processing of the human auditory system. Since humans are the best interpreters of speech, it seems reasonable to assume that the auditory processing may enhance important events in the acoustic signal, while de-emphasizing others. Some researchers have developed algorithms and displays which attempt to model the auditory processing (Searle et al., 1980; Lyon, 1984; Ghitza, 1988; Seneff, 1988; Shamma, 1988). With the popularity of digital computers the spectrogram has become more versatile, and some of its drawbacks have been addressed. Today, many laboratories have developed facilities for producing digital spectrograms, with quality comparable to the analog spectrograms. An advantage of digital processing is that it is easy to modify the analysis and display parameters. Kay Elemetrics Corp. has a commercially available digital spectrograph machine, the Kay DSP Sonograph. The DSP Sonograph also provides the capability to display other parameters such as the waveform and energy envelope, linear prediction analysis, and spectral slices at a given point in time. A spectrogram of an unknown utterance, produced using a Voice-Print, model 4691A, is shown in Figure 1.1(a). Part (b) of Figure 1.1 shows the same utterance produced by the DSP Sonograph, model 5500. Figure 1.2 shows a typical spectrographic display used at MIT, and in this thesis, for the same utterance. It was produced using the software tool *Spire* (Shipman, 1982; Cyphers, 1985). The spectrogram was computed by taking the



(a)



(b)

Figure 1.1: Example spectrogram (a) produced by the Voiceprint, (b) produced by the Kay DSP Sonograph.

short-time Fourier transform (STFT) of the speech signal

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m} \tag{1.1}$$

where $x[m]$ is the digitized speech signal, n is the time index, and w is a Hamming window of 6.7 ms. The STFT is computed every millisecond and sampled at 128 frequency points over the frequency range 0 to 8 kHz. The amplitude is then nonlinearly mapped into a 25 dB grey scale. The spectrogram is augmented by three parameters: low frequency energy (LFE), total energy (TE) and center-clipped zero crossing rate (ZCR), along with the original waveform display. These parameters are useful to the spectrogram reader in identifying phonemes, particularly in regions where the acoustic energy is weak. For example, some weak fricatives are not apparent on the spectrogram and can only be postulated by the presence of a high ZCR. Researchers may augment the spectrogram with other parameters. Vaissiere (1983) has found that the fundamental frequency contour aids in interpreting spectrograms of French sentences.

1.2 Spectrogram reading

Some humans have learned to interpret the visual acoustic patterns in the spectrogram so as to determine the identity of the spoken phonemes or words, a process known as spectrogram reading. In addition to providing a convenient mechanism for studying acoustic-phonetics (the relationship between phonemes and their acoustic correlates), spectrogram reading provides an opportunity to separate the acoustic characteristics of sounds from other sources of information, such as lexical, syntactic and semantic. It is difficult to assess the role of the different knowledge sources used by listeners interpreting continuous speech. That lexical, semantic and pragmatic knowledge are important is demonstrated by the ability of listeners to understand speech even under distortion. Humans are also capable of decoding the intended message in the presence of speech errors (Nickerson and Huggins, 1977). The importance of language-specific knowledge was demonstrated by experiments in which phoneticians were asked to transcribe utterances from both familiar and unfamiliar languages (Shockey and Reddy, 1975). The phoneticians were typically less consistent at transcribing unfamiliar languages, suggesting that language-specific knowledge is important for phonetic decoding.

It can be argued that in reading spectrograms one may be able to use fewer sources of knowledge than one can in listening. Spectrogram readers may be able to rely on their knowledge of the acoustic characteristics of speech sounds, how these characteristics change due to coarticulation, and on phonotactics, the allowable sequences of phonemes in the language. It appears that the spoken phonemes may be labeled in the spectrogram without considering word hypotheses. The claim is not that one cannot or should not try to read words or phrases directly in the spectrogram, but that it is possible to interpret the spectrogram without reading the words. The aim of Potter, Kopp, and Kopp (1947) was to assess the feasibility of communicating via spectrograms. Other researchers have also investigated reading words or syllables directly (House et al., 1968; Greene et al., 1984). This thesis work has focused on relating the visual acoustic patterns in the wide-band spectrogram to the underlying phonetic representation.

The earliest research in spectrogram reading was undertaken by Potter, Kopp and Kopp at Bell Laboratories in 1943.[1] As noted in the book *Visible Speech* (1947) they first presented evidence of readability:

> Different words have a different appearance, an essential requirement if they are to be told apart. But the same words spoken by different individuals have a similar appearance, also an essential requirement if the symbols are to be of practical use. [p.5]

The purpose of their research was to develop a speech communication aid for the deaf. Spectrogram reading was studied along with phonetic principles and the relationship of articulatory movements to speech patterns. The studies were reported in *Visible Speech.* The book provides a comprehensive summary of the acoustic/visual properties of speech sounds, and to date remains the only published book on this topic. Rothenberg (1963) wrote a manual for interpreting spectrograms and Fant (1968) provides a guide to phonetically interpreting spectrograms.

Much of the pioneering work in acoustic-phonetics (Lehiste, 1967) focused on small units of speech, typically simple syllables and words. The analysis of consonant-vowel-consonant (CVC) or VCV sequences provides valuable insight into the canonical acoustic

[1] A completely independent study is described in a book by Solzhenitsyn, *The First Circle* (1968). In this book a scientific prisoner, Lev Rubin, learned to read speech patterns in a secret project under Stalin. An example of identifying an unknown speech signal is given on page 189. The extent to which this account is true is unknown.

characteristics of speech sounds. These studies also defined some of the acoustic correlates of distinctive features (Jacobson et al., 1952). For example, the distinction between voiced and voiceless sounds in English is often cued primarily by duration rather than vocal-fold vibration (Denes, 1955; Lisker, 1957, 1978; Klatt, 1976; Umeda, 1975, 1977). The acoustic characteristics of speech sounds can also be related to their articulation. Formant locations for vowels and the spectral energy present in consonants can be predicted by acoustic-tube models of vocal tract configurations (Fant, 1960).

Despite the early work of Potter et al., and the role of spectrograms in speech analysis, the prevailing opinion was that speech spectrograms were extremely difficult to read (Liberman et al. 1967,1968). While Fant (1962) argued for the utility of reading speech spectrograms, he also noted that no researchers claimed to be able to read them fluently. A common assumption was that the coarticulation between sounds was such that it would obscure the identity of individual phonemes. Some researchers believed that the acoustic signal, by itself, does not provide enough constraint to uniquely decode the utterance, but that higher-level constraints obtained from syntax and semantics must be used (Newell et al., 1971; Reddy, 1976). Studying spectrograms of continuous speech may help us to better understand acoustic-phonetics and the phonological variation found in continuous speech. For example, it is well known that the acoustic characteristics of the words "did" and "you" spoken in isolation are quite different from their common pronunciation as [dɪ̆ju] in fluent speech. Only by directly studying the acoustic characteristics of fluent speech can such phonological variation be understood.

Spectrogram reading has contributed to our understanding of acoustic-phonetics and indirectly contributed to speech synthesis and recognition. Real-time spectrograms and other devices have also been used to correct speech production problems in hearing-impaired subjects (Stewart et al., 1976; Houde and Braeges, 1983). Spectrogram reading has also had two direct applications. Reading spectrograms has been proposed as an alternative method of communication for the deaf, and as a potential aid for the hearing impaired (Potter et al., 1947; House et al., 1968, Nickerson, 1978; Cole and Zue, 1980). Recently researchers have attempted to build automatic speech recognition systems that explicitly model spectrogram reading (Johanssen et al., 1983; Carbonell et al., 1984; Johnson et al., 1984; Stern, 1986; Stern et al., 1986).

1.3 An example of interpreting a spectrogram

Reading spectrograms involves the application of a variety of constraints to the identitication problem. These include knowledge of the acoustic correlates of speech sounds and their contextual variation, and phonotactic constraints. The skill also requires the ability to integrate multiple cues and to rely on secondary cues when the primary ones are not present.

Protocol analysis of the spectrogram reading process (Cole and Zue, 1980) shows there to be two stages, roughly corresponding to segmentation and labeling. Segmenting the speech involves placing boundaries to mark acoustic change. Boundaries are usually marked where there is a large spectral discontinuity. However, often the only cue to a vowel-semivowel transition is the amount of gradual formant motion. Other segment boundaries, such as for geminate consonants, may be cued only by duration. Experienced spectrogram readers often do not explicitly mark boundaries, but rather implicitly denote them via the labeling. Generally the easy segments, those whose spectral patterns are distinct and relatively context invariant, are labeled first. Then, with successive revisions, incorporating the local context and finer acoustic cues, the remaining segments are labeled. Phonotactic constraints may also aid in the process. Although there may be feedback in the process (a partial identification of the segment may help in further segmentation), often the stages may be separated.

In order to illustrate the process of spectrogram reading and to relate some of the properties of speech sounds to their visual patterns in the speech spectrogram, I will walk through the identification of the phonemes in the spectrogram in Figure 1.2. For ease of discussion, the phonemes are interpreted from left-to-right. Throughout the example the acoustic characteristics of the speech sounds are related to their articulation. For a comprehensive analysis of the relationships see Fant (1960) and Flanagan (1972). Spectrogram reading brings together information from a variety of sources in forming phonetic labels. I intend for this illustration to give the reader a flavor for the process; this example is not meant to be comprehensive.

The utterance begins with a stop release at time t=0.05 sec. The voice-onset time (VOT) of the stop is quite long, about 90 ms. The long VOT and the presence of aspiration indicate that the stop is voiceless. Stops are produced by forming a complete

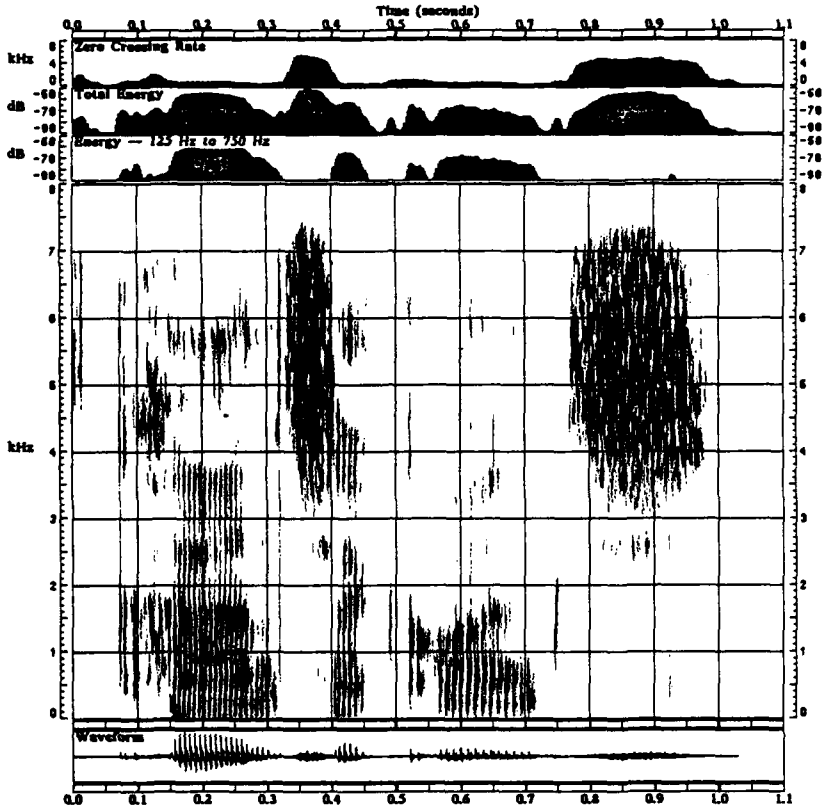


Figure 1.2: Example spectrogram produced using *Spire*. The display also includes low frequency energy (Energy – 125 Hz to 750 Hz), total energy, and zero crossing rate contours. The waveform is shown below the spectrogram.

constriction in the vocal tract, and abruptly releasing the constriction. Only the cavities in front of the constriction are initially excited; thus the spectral characteristics of the release provide information about the place of articulation of the stop (Fant, 1960). The spectral distribution of energy at the release has two major concentrations of energy. The lower concentration is centered at about 1600 Hz, approximately the same frequency as the second formant (F_2) of the next vowel. The higher concentration is at almost three times the lower frequency. This bimodal frequency distribution is typical of velar stops, where the energy concentrations correspond to the first two resonances of the quarter-wavelength acoustic cavity in front of the constriction. Thus, the first segment is a /k/.

The next segment (from t=0.14 sec to t=0.25 sec) is a vowel with a high F_1 and a low F_2. Based on the formant locations the vowel has the distinctive features [+ low] and

[+ back] and is probably an /ɑ/ or /ɔ/ (Jacobson et al., 1952).

Following the vowel is a nasal (from t=0.25 sec to t=0.3 sec). The presence of the nasal is primarily indicated by the abrupt spectral change at the end of the vowel: the disappearance of the higher formants, and the appearance of the low nasal resonance, at about 250 Hz (Fujimura, 1962; Mermelstein, 1977). In fact, the nasal resonance actually extends back into the preceding vowel: this is evidence of nasalization of the vowel (Fujimura, 1960). The place of articulation of the nasal is not obvious. In this case the candidates are ordered by the lack, rather than the presence of acoustic evidence. The third formant is rising from the vowel into the nasal, indicating that the nasal is probably not labial or velar. However, if the nasal is alveolar, then F_2 should to rise towards a locus near 1800 Hz (Delattre et al., 1955; Halle et al., 1957), but there does not seem to be much motion in F_2. If the nasal is labial, F_2 should be falling into the nasal and there may be a lowering of the spectral energy distribution at the beginning of the following fricative. Perturbation theory (Fant, 1960) predicts both of these effects as a consequ nce of forming a constriction at the lips. To label the segment more precisely than simply "nasal," I would rank the nasals in the order /n/,/ŋ/,/m/.

Following the nasal is a strident fricative, indicated by the strong noise-like energy at high frequencies. The high total energy and zero crossing rate provide supporting evidence. Strident fricatives are produced by forming a narrow constriction with the tongue in the oral part of the vocal tract such that turbulent noise is generated at an obstruction anterior to the constriction. The noise source excites the cavity in front of the constriction. (The resonant frequency of the cavity is inversely proportional to its length.) The resonances of the cavities behind the constriction are cancelled by zeros (Fant, 1960). In this case, the energy is primarily above 4 kHz, indicating that the fricative is alveolar, and therefore an /s/ or a /z/. The duration of the fricative is about 80 ms, which is not particularly short or long. The lack of voicing cues, such as vertical striations in the noise or periodicity in the waveform, tend to favor /s/ as the top choice.

The next segment is a short vowel; it is only about five pitch periods long, suggesting that it is [- tense] and not stressed. The first and second formants are further apart in frequency than the first vowel, indicating that this vowel is more fronted, an /ɛ/ or an /ɪ/.

The second and third formants come together at the end of the vowel (t= 0.43) in what

is referred to as a velar pinch. This formant motion is typical in a front vowel next to a velar. Following the vowel is an interval of silence, corresponding to a stop closure. The release of the stop is at t=0.5 sec. The characteristics of the stop release, compact in frequency and located near F_2 of the next sonorant region, support the evidence in the preceding vowel that the place of articulation is velar. The stop also has what is known as a double burst in time, another indicator for a velar place of articulation (Fischer-Jørgenson, 1954; Keating et al., 1980). In fact, the first stop in the utterance also has a double (maybe even triple) burst. The cues for voicing of the stop are mixed. The VOT of the stop is about 50 ms, compatible with both a /g/ and a /k/. Conflicting are the lack of prevoicing in the closure (which would support a voiced stop) and the lack of aspiration in the release (which would favor a voiceless stop). The lack of aspiration can also be seen by comparing the zero crossing rate in the two stops. The stop is either a /g/ or an unaspirated /k/.

The stop release at t=0.5 sec is lower in frequency than observed for the first velar stop. This is because the next segment is rounded, a /w/. The presence of the /w/ is indicated by the low F_1 and F_2 at the beginning of the voiced region, and the rising formant motion into the vowel. (An /l/ may be a second choice, as /l/ also has a low first and second formant. A variety of cues lead me to favor /w/. These include the especially low frequency of F_2, the low frequency location of the burst, and the lack of higher frequency energy in the release often present with /l/.) Stops in semivowel clusters typically have longer VOT values than singleton stops (Klatt, 1975; Zue, 1976), suggesting that this stop is a /g/. However, the cluster /gw/ is relatively rare in English and the total energy contour indicates that the final syllable of the utterance is less stressed than the initial one. Thus an unstressed, unaspirated /kw/ cluster is also possible.

The acoustic characteristics of the final vocalic portion are not particularly clear. The first formant is neither high nor low in frequency and the second formant location is affected by the preceding /w/. At its midpoint, roughly t=0.62 sec, the vowel looks to be relatively neutral, probably /ʌ/ or /ɛ/. The end of the vowel appears nasalized (the bandwidth of the first formant is large) and there is a nasal murmur from t=0.65 sec to t=0.7 sec. The nasal resonance also extends back into the preceding vowel. The place of articulation of the nasal is difficult to determine as the formants in the preceding vowel fade away before providing any clear indications. However, F_2 in the preceding segment is rising more than expected if the nasal were labial, and less than would be expected for

a velar (compare the F_2 motion to the vowel at t=0.4 sec). F_2 may be heading to a locus near 1800 Hz, indicating alveolar. There is a period of silence lasting approximately 50 ms followed by an /s/ at t=0.75 sec. A /z/ is ruled out because the nasal murmur should be longer if the nasal were followed by a voiced consonant in the same syllable (Malécot, 1960; Raphael et al., 1975; Zue and Sia, 1982). The silence may be due to articulatory timing or may be a stop closure. If it is a stop, it is homorganic (has the same place of articulation as) with the nasal. The lack of spectral change in the /s/ suggests that the preceding nasal is most likely an /n/.

The phoneme string thus proposed is

k	ɑ	n	s	ɛ	k	w	ʌ	n—(t)	s
	ɔ		z	ɪ	g		ɛ		
		m				l		ŋ—(k)	
		ŋ						m—(p)	

where, being conservative, the phonemes below the dashed line are less likely, but have not been definitively ruled out. From this transcription it is easy to obtain the word proposal "consequence." In fact, in a 20,000 word lexicon (Webster, 1964) it is the only word matching the transcription.

I have used this example to demonstrate that the process of spectrogram reading entails identifying acoustic characteristics of phones and using a combination of constraints. Typically a "broad class" phoneme proposal, such as nasal, stop or fricative, is refined using more detailed evidence. Some segments, such as the /s/ and /k/, are identified by recognition of their canonical characteristics. An example of contextual variation is illustrated by the two /k/'s in the utterance. Although both precede a sonorant that is [+ back], the second /k/ has a somewhat lower burst frequency since it is also rounded. The two /k/'s also exhibit differences due to stress.

1.4 Summary of spectrogram reading experiments

After the pioneering work in 1947, spectrogram reading was not actively pursued until the early 1970's, spurred by the interest in automatic speech recognition. Around this time

exploratory studies were performed (Klatt and Stevens, 1973; Lindblom and Svenssen, 1973; Svensson, 1974), with somewhat discouraging results. In a series of experiments in 1978 and 1979 (Zue and Cole, 1979; Cole et al., 1980; Cole and Zue, 1980), Zue demonstrated that spectrograms of continuous speech could be phonetically labeled with accuracy better than 80%. A summary of these and subsequent spectrogram reading experiments is given in Table 1.1. Blanks are left in the table when the relevant data were not given in the reference. While the spectrogram reading experience of many of the subjects was unspecified, most subjects were researchers in speech or linguistics and familiar with acoustic phonetics. The accuracy reported in the table is for the top choice phoneme unless otherwise indicated.

As can be seen in Table 1.1 there have been a variety of spectrogram reading experiments. Some of the experiments addressed the ability of subjects to read words or syllables directly in the spectrogram (Potter et al., 1947; House et al., 1968; Pisoni et al., 1983; Greene et al, 1984; Daly, 1987). Others attempted to assess the ability to phonetically label the spectrogram (Klatt and Stevens, 1973; Svenssen, 1974; Cole et al., 1980, Johnson et al., 1984; Lonchamp et al., 1985). The subjects' performance at phonetic labeling ranges from a low of near 30% to a high of around 80%. Some of this variation may be attributed to the test conditions. For example, the subjects in the Svensson (1974) study were instructed to provide only one label per segment; in other experiments multiple labels were permitted. In order to maximize the likelihood that the labeling was based on acoustic-phonetic evidence and to minimize the possibility of hypothesizing words, Klatt and Stevens slid a 300 ms window across the sentence in a single left-to-right pass. However, the window also prevented the readers from using utterance-based "normalization," such as for fricative energy or formant locations. In the other studies, the subjects were able to see the entire spectrogram at once. The conditions of the experiments vary so much with regard to the test data, test conditions, and subject experience that it is difficult to compare the results. With such a range of experiments and results, it is no wonder that the possibility of reading spectrograms has been questioned (Liberman et al., 1968).

The experiments of Cole et al. (1980) were the first to indicate that a highly trained expert could phonetically label a spectrogram of an unknown utterance with an accuracy better than 80%. The labels produced by Zue were compared to the phonetic transcriptions of three trained phoneticians: one of the labels provided by Zue (at most three

Table 1.1: Comparison of previous spectrogram reading experiments.

Author	date	Number of subjects	Experience/ training	Number/Sex of talkers	Type of utterance	Number of segments	Accuracy	Comments
Potter et al.	1947	5	90h	5 f	words	300 words		Visual Speech translator
		1 (deaf)	200h		words	780 words		no measure of performance accuracy
House et al.	1968	8			8 /t/-V's		55-80% Vs	Visual Speech translator
Klatt&Stevens	1973	2	researchers	5 m	19 sentences	658	33%,73% partial	left-to-right analysis, 300 msec window
Lindblom&Svensson	1973	1	experienced	1 m	9 sentences	237	99%,78% sentence	Swedish, with prosody and instructions
		7	researchers/	1 m	9 sentences	1629	~50%,27% words	" " " "
		6	students	1m	9 sentences	1422	~40%,7% words	no prosody
Svensson	1974	14	"	1 m	9 sentences		38% (range 22-51%)	Swedish anomolous sentences
Kuhn&McGuire	1974	2	researchers	1 m	VCV	432	83% Cs	improved from 78% to 90% with practice
Sendlt	1978	1	30-40h	1 m	40 sentences	1183	84%, 60% word	improved from 80% to 89% with practice
Cole et al.	1980	1	2500-3000h	2 m	23 sentences	499	85% top 3	performance compared with phoneticians
					46 words	173	93% top 3	" " "
Cole&Zue	1980	5	10 hours	1 m	5 sentences	104	51%, 83% top 3	group analysis,general instructions
		1	2500-3000h	1 m	11 sentences	170	81% top 3	how last can Zue label
					25 words	97	82%, 93% top 2	
Johnson et al.	1984	1	phonetician	1 m, 1 f	60 sentences		80%,60% sentence,68% word	number of labels not specified
Pisoni et al.	1983	10	27h	1 m	50 words		40%,68% manner	subjects took 1 week spectrogram reading course
Greene et al.	1984	8	22h	1 m	50 words		95% training words	subjects trained to read words
		8	22h		50 words		91% new tokens of words	same speaker
		8	22h	1 m, 1 f	50 words		76% new talkers	
Lonchamp	1985	1	researcher	5 m	50 sentences	1172	70% Cs, 80% top3 Cs,73%V	French, vowel(V) accuracy reported for features
Stern et al.	1986	1	experienced	1 m	100 sentences	887	95% in trills	French sentences using13 phonemes
Daly	1987	10	> 50 hours	2 m, 2 f	26 letters	1040 letters	89% letter,92% top 2	isolated letters
		6	> 300 hours	10 m, 10 f	1000 strings	5601 letters	91% letter,95% top 2	strings of letters

choices were supplied) agreed with at least one of the transcribers over 85% of the time. Zue's performance is particularly encouraging in light of the observation that the agreement as to the correct answer among the three phoneticians was also 85%. However, a question remained as to whether or not the skill of spectrogram reading could be taught. Would all speech researchers interested in spectrogram reading have to invest 2000 hours, as Zue did, to become proficient at the task? Cole and Zue (1980) report an experiment in which Zue, as part of a course on speech production and perception at Boston University in 1978, attempted to teach a group of five graduate students how to read spectrograms (see Table 1.1). A year later, Seneff (1979), serving as her own subject, conducted an experiment in which she labeled spectrograms of 49 sentences. After each spectrogram was read, Seneff discussed the spectrogram with Zue. Seneff was encouraged that her performance, with regard to both accuracy and speed, improved rapidly. More recently, a number of spectrogram reading courses have been taught and several researchers have become proficient at this task. The growing interest in spectrogram reading is apparent by the popularity of the *Speech Spectrogram Reading* courses taught at MIT over the last five years. The success of these courses provides evidence that the knowledge used in spectrogram reading can be transferred.

The results of some of the more recent spectrogram reading experiments are quite encouraging (Cole and Zue, 1980; Johnson, 1984; Lonchamp, 1985). These results suggest that the accuracies with which spectrograms in different languages are phonetically labeled, by a subject familiar with that language, may be comparable. The studies indicate that trained spectrogram readers can phonetically label an unknown utterance with better accuracy than existing phonetic speech recognition systems (Klatt, 1977). However, one should be cautious in interpreting these results, as the tests were quite limited and and the conditions varied. The data on which the subjects were tested ranged from simple CV-syllables to continuously spoken sentences. The amount of test data was generally small, as was the number of subjects. The limited testing is not surprising, as the evaluation is rather time-consuming and often requires highly trained subjects. The experience of the subjects also varied greatly, from naive to experienced. In addition, in almost all of the studies, speech from only a small number of talkers, typically 1 to 5 male talkers, was used. (The talkers also tended to be speech researchers at the laboratory where the experiment was conducted.) The small scale of the experiments and the lack of consistency among them indicates the need for a more extensive evaluation.

1.5 Scope of the thesis

While human listeners are the best speech recognizers, some human viewers have learned the skill of interpreting the patterns present in spectrograms to determine the identity of the spoken phonemes. The phonetic transcription thus obtained is as good or better than can presently be achieved by automatic speech recognition phonetic front ends (Klatt, 1977; Zue and Cole, 1979; Cole et al., 1980). Researchers have learned much about acoustic-phonetics from extensive studies of speech spectrograms and have been incorporating knowledge and features derived from the study of spectrograms in speech recognition systems (see, for example, Cole et al., 1982; Demichelis et al., 1983; Glass, 1984; Chen, 1985; Espy-Wilson, 1987). Some researchers have attempted to develop expert systems which attempt to mimic spectrogram reading (Johanssen et al., 1983; Johnson et al., 1984; Carbonell et al., 1986; Stern, 1986; Stern et al., 1986).

It is evident from the spectrogram reading experiments that the acoustic signal is rich in phonetic information. The phonetic segments in the utterance are located and labeled from the visual representation of the speech signal. Several sources of knowledge are used to interpret a spectrogram. These include knowledge of the characteristic visual patterns of speech sounds, how these patterns are modified due to coarticulation, and phonotactic constraints. Many of the observed acoustic correlates of speech sounds can be predicted by articulatory models and some of the contextual variations can be explained using perturbation theory and simple acoustic-tube models (Fant, 1960). In this thesis I am concerned with relating the visual patterns in the spectrogram to phonetic units without the use of higher-level knowledge, such as lexical, syntactic or semantic knowledge.

While the results of previous spectrogram reading experiments are quite encouraging it must be kept in mind that the evaluations were on fairly small test sets, spoken by a small number of talkers. It is not apparent from the reported experiments whether or not accurate phonetic labeling of speech from many different talkers can be obtained. Thus, one of the aims of this thesis has been to systematically evaluate experienced spectrogram readers on speech from a large number of speakers and in a variety of local phonemic contexts. The results of spectrogram reading experiments on a task that does not permit the use of higher-level knowledge are presented in Chapter 4.

How should the ability of humans to phonetically label spectrograms be assessed? A logical comparison is with trained phoneticians, as reported by Cole et al. (1980). Two

problems associated with such an approach, namely, inter-transcriber consistency and the use of higher-level knowledge, were discussed in the previous section. I have opted to evaluate naive listeners on the same task as the spectrogram readers. The listeners are "naive" in the sense that they are not trained phoneticians, but being speaker/hearers they have had years of experience at listening. The listening experiments serve both as a baseline performance measure and to determine whether or not factors thought to be important in spectrogram reading are also important to listeners. (Spectrogram reading alone does not indicate whether or not the acoustic patterns and rules used by spectrogram readers bear any correspondence to the information used by listeners.) In order to minimize the use of higher-level knowledge, listeners heard portions of speech extracted from continuous sentences. The listening experiments, presented in Chapter 3, vary factors such as stress, syllable position and phonetic context.

The evidence, obtained from both spectrogram reading experiments and from teaching spectrogram reading, indicates that the process can be modeled with a set of rules. Formalizing spectrogram reading entails refining the language (terminology) that is used to describe acoustic events on the spectrogram, and selecting a set of relevant acoustic events that can be used to distinguish phones. Rules which combine these acoustic attributes into phones must also be developed. The rules need to account for contextual variation (coarticulation), and partial and/or conflicting evidence, and to be able to propose multiple hypotheses. One way to assess how well the knowledge used by experts has been captured in the rules is by embedding the rules in a computer program. The knowledge may be explicitly incorporated in a knowledge-based system. The degree to which the knowledge has been formalized can be judged by the performance of the system, the types of errors made by the system, and the reasoning used.

Building a system to label entire, unrestricted utterances is beyond the scope of this thesis. I hope, however, to take a step in that direction. The specific task investigated in this thesis is the identification of stop consonants extracted from continuous speech. The stops occur in a variety of contexts, including both syllable-initial and syllable-non-initial position, and in clusters with nasals, semivowels, and fricatives. The contexts were chosen to test the importance of knowledge sources thought to be used in spectrogram reading. A partial segmentation of the speech is provided. Restricting the information to the segment to be identified and its immediate neighbors greatly reduces the complexity of the problem while retaining much of the contextual influences in American English.

The remainder of this thesis is organized as follows. In Chapter 2 the design of the listening and spectrogram reading experiments is discussed. Examples of relevant acoustic characteristics are also provided. Chapters 3 and 4 present the results of the listening and spectrogram reading experiments, respectively. The acoustic attributes, rules, and knowledge representation used in the knowledge-based system are presented in Chapter 5. Included in Chapter 5 is an evaluation of the system. A final discussion and suggestions for future research are given in Chapter 6.

Chapter 2

Task and Database Descriptions

This chapter provides a description of the tasks used to evaluate human listeners, spectrogram readers, and the knowledge-based implementation. Factors such as stress, syllable position and phonetic context were varied in order to determine their effects on stop identification. The test tokens were extracted from continuous speech spoken by many talkers. The next section provides an overview of the organization of the experiments. More detailed discussions of each task are given in section 2.2. The final section specifies details of the token selection and distributional properties of the test data.

2.1 Organization of the experiments

The experiments described in Chapters 3 and 4 assessed the subjects' ability to identify stop consonants presented in only their immediately surrounding phonetic context. The tokens, extracted from continuous speech, consisted of a stop or a two-consonant sequence containing a stop, and a single vowel on each side. The experiments were designed to explore whether sufficient acoustic information was present in the extracted waveforms to identify the stops. There are several reasons why stop consonant identification was selected for this experiment. As a class of sounds the stop consonants have been extensively studied. Their articulation is complicated, consisting of dynamic characteristics which vary depending on context (e.g., Fant, 1960; Fant, 1973). Stops are also among the most frequently occurring sounds in English, appearing both alone and in a variety of consonant clusters. They account for roughly 20% of all phonemes (Denes, 1963). A variety of studies on the perception of stops in isolated CV syllables have shown an identification performance in the range of 97% to 99%(Nye and Gaitenby, 1973; Pisoni and

Hunnicutt, 1980). In syllable-final position the identification rates drop by 2-5%(House et al., 1965; Nye and Gaitenby, 1973).

Subjects identified stop consonants in five different contexts:

Task 1: syllable-initial singleton stops
Task 2: syllable-initial stops preceded by /s/ or /z/
Task 3: syllable-initial stops in semivowel clusters and the affricates
Task 4: non-syllable-initial singleton stops
Task 5: non-syllable-initial stops in homorganic nasal clusters

The first task assesses the subjects' ability to identify singleton stop consonants in syllable-initial position. After establishing this baseline performance, the effects of intervening consonants and syllable position on the subjects' decision can be determined. Acoustic studies have shown that the acoustic characteristics of stops in syllable-initial consonant clusters change from the canonical characteristics of singleton stops (Lehiste, 1962; Zue, 1976). The remaining tasks evaluate the subjects' ability to identify stop consonants in clusters with other consonants and in non-syllable-initial position.

The five tasks were combined into experiments as shown in Figure 2.1. Experiment I compared tasks 1 and 2 assessing the effect of alveolar strong fricatives on the identification of syllable-initial stops. In Experiment II, comparing tasks 1 and 3, the question raised was whether the semivowels alter the identification of the stop consonants. Experiment III examined the extent to which syllable position affects stop identitication. Nye and Gaitenby (1973) found syllable-final stops to be less well articulated than syllable-initial stops. Experiment IV investigated the influence of homorganic nasals on the identification of non-initial stops. The next section describes the tasks in more detail.

2.2 Description of the tasks

In the first task, subjects identified syllable-initial singleton stop consonants. Syllable-initial stops generally are well-articulated and exhibit their canonical characteristics (Halle et al., 1957; Fant, 1973; Zue, 1976). In English, a primary cue for voicing in syllable-initial singleton stops is the voice-onset-time (VOT) (Lisker and Abramson, 1964). Other cues include the presence/absence of aspiration after the burst and prevoicing during the closure interval. Lisker (1978) catalogues 16 acoustic features that may

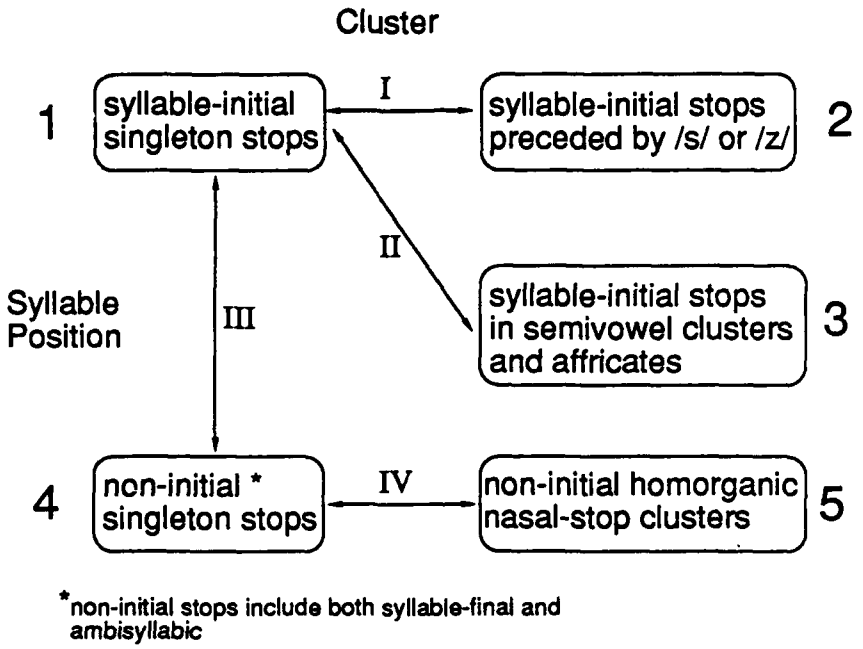
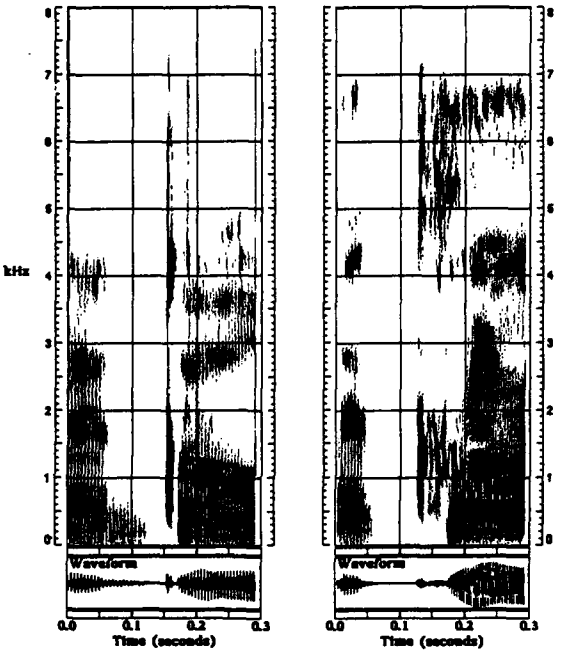


Figure 2.1: Experimental design. Task 1: syllable-initial singleton stops; Task 2: alveolar strong fricatives (/s, z/) preceding syllable-initial stops where the fricative may or may not be in the same syllable as the stop; Task 3: syllable-initial stops in clusters with semivowels /l,r,w/ and the affricates, /ʧ,ʤ/; Task 4: non-syllable-initial singleton stops; Task 5: non-syllable-initial nasal-stop sequences. The roman numerals I,II,III, and IV denote the experiment number.

cue the voicing distinction. The acoustic characteristics of the stop release provide information about the place of articulation as do the formant motions into the surrounding vowels. This task set a baseline performance measure for the ensuing tasks, and explored whether or not the immediate phonetic context was sufficient for identification of the stop.



/əgo/ /əko/

Figure 2.2: Spectrograms of /əgo/ and /əko/.

Spectrograms of a syllable-initial /g/ and /k/ are shown in Figure 2.2. The VOT of the /g/ (about 20 ms) is shorter than the VOT of the /k/ (about 60 ms). The prevoicing throughout closure of the /g/ and the aspiration following the release of the /k/ provide additional evidence for voicing. The spectral characteristics of the release of the stops in Figure 2.2 are quite similar and typical of a velar place of articulation.

In task 2, an alveolar strong fricative (/s/ or /z/) preceded a syllable-initial stop, where the fricative may or may not have been in the same syllable as the stop. The presence of the fricative potentially affects the identification of both the place and voicing of the stop. Since a fricative precedes the stop, the formant transitions out of the preceding vowel should always indicate an alveolar place of articulation for the fricative instead of

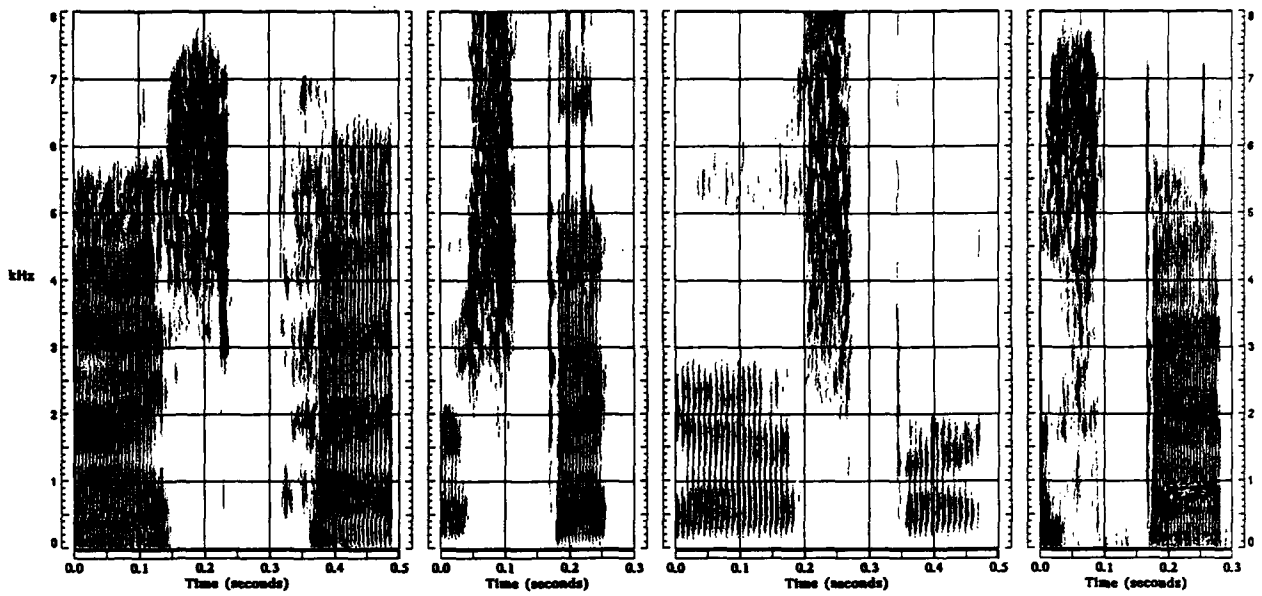


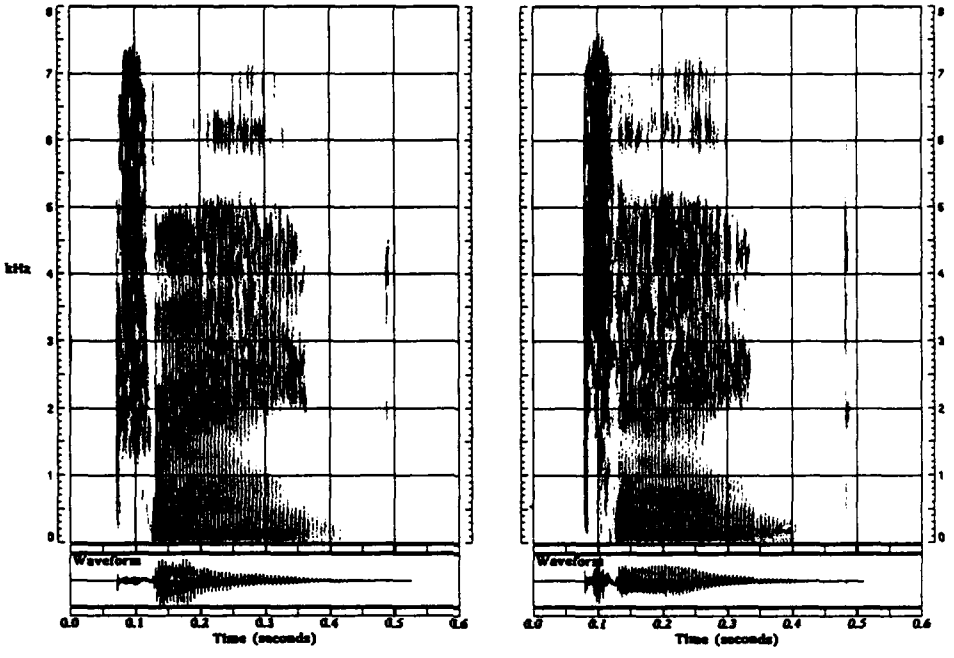
Figure 2.3: Spectrograms of /ɑʸs–pɛ/,/ə–spɛ/,/æs–bɝ/ and /əz–bæ/.

indicating the place of articulation of the stop. However, cues to the place of articulation of the stop may be present at the end of the fricative. An example of one such cue can be seen in the leftmost spectrogram in Figure 2.3. The lower frequency limit of energy at the end of the /s/ is falling into the stop. This pattern is called a labial tail, and occurs because the lips move to form the stop closure while the fricative is still being produced.[1] The voiceless stops (/p,t,k/) are typically unaspirated when they are in a cluster with an /s/ and have a shorter VOT (Davidsen-Nielsen, 1974; Klatt, 1975; Zue, 1976). The lack of aspiration and reduced VOT may lead to errors in the identification of voicing if subjects are unable to determine that the stop is in an /s/-cluster. The remaining spectrograms in Figure 2.3 illustrate the similarity among an /sp/-cluster and a /b/ preceded by an /s/ and a /z/.

Phonotactic constraints may also be applied in this task. For example, if the subject could identify the fricative as a /z/, then the subject knew that there must be a syllable boundary before the stop, and that syllable-initial voicing cues should be used. Since the

[1]The same pattern is also common preceding phonemes that are rounded. Perturbation theory (Fant, 1960) predicts the lowering of resonant frequencies due to lengthening the front cavity by protruding the lips or as a consequence of forming a constriction at the lips.

identity of the fricative may have influenced the identification of the stop, subjects were also asked to identify the fricative as either an /s/ or a /z/.

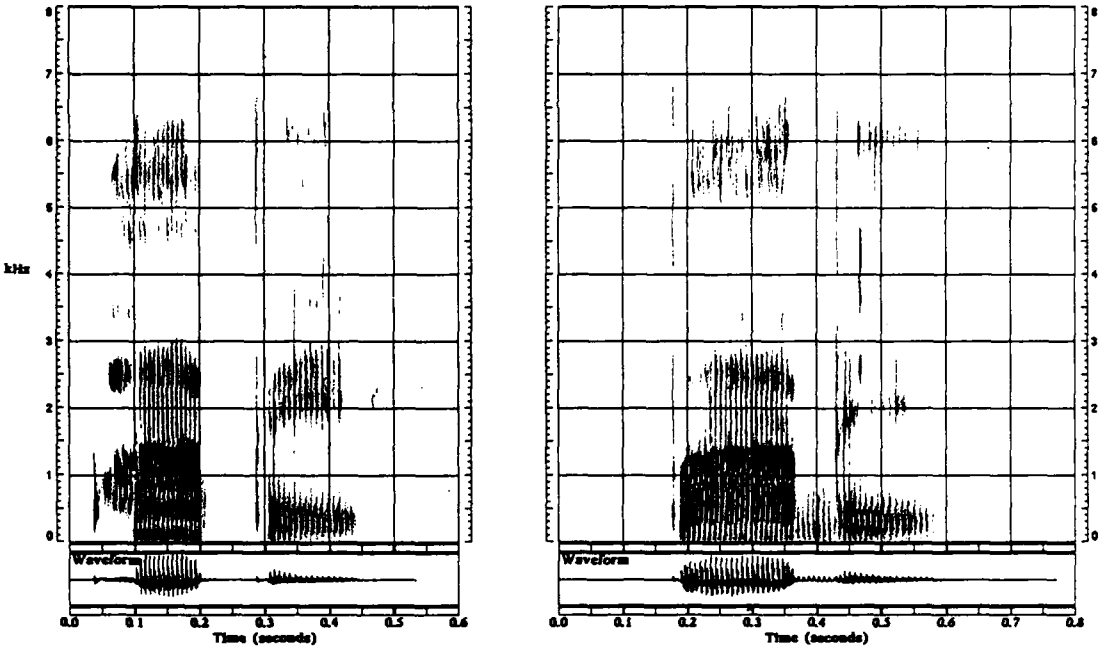


"drain" "Jane"

Figure 2.4: Spectrograms of "drain" and "Jane."

The stimuli in task 3 consisted of tokens of syllable-initial stop-semivowel clusters and of affricates, /ʧ,ʤ/. This task investigated the effect of the semivowels /l,r,w/ on stop consonant identification. Earlier acoustic studies (Lehiste, 1962; Klatt, 1975; Zue, 1976) have shown that semivowels affect the acoustic characteristics of neighboring sounds. In particular, semivowels tend to strengthen and lengthen the release of a stop and change its spectral characteristics. There is often a longer period of frication noise than observed for singleton stops which may cause voiced stops to be mistakenly identified as voiceless. The affricates were included in order to determine if the increased frication present in /dr/ and /tr/ clusters was sufficient to make them confusable with affricates. Figure 2.4 illustrates the acoustic similarity of the words "drain" and "Jane." Phonotactic constraints can also be applied in this task, as certain stop-semivowel combinations (such as a syllable-initial /tl/) are not permissible.[2]

[2]While theoretically such sequences cannot occur, in reality they sometimes do. For example, the reduced vowel in "Toledo" can be deleted, leaving behind the sequence [tl]. This is a fairly rare occurrence and is therefore not considered here.



"poppy" "bobby"

Figure 2.5: Spectrograms of "poppy" and "bobby."

Some researchers have argued for the syllable as a unit of representation in phonology (for example, Kahn, 1976). As such, syllable position is expected to play a role in speech production and perception. Task 4 assessed the subject's ability to identify singleton stops in non-syllable-initial position. Non-syllable-initial refers to both syllable-final stops and ambisyllabic[3] stops. Non-syllable-initial stops are more difficult to identify than syllable-initial stops, since they often do not exhibit as robust a release. Voiceless stops in non-initial position frequently are unaspirated, making the determination of voicing much harder. Although syllable-final stops are often not released, only those transcribed as having both a closure interval and a release were used as stimuli.

Figure 2.5 shows spectrograms of the words "poppy" and "bobby." The initial stop in each word exhibits its typical, syllable-initial, prestressed characteristics. The spectral amplitude of the release is weak in relation to the vowel, with the energy distributed

[3]According to Kahn (1976) ambisyllabic consonants are those shared by two syllables. They occur in instances where placement of a syllable boundary is arbitrary: "it makes sense to speak of *hammer* as consisting of two syllables even though there is no neat break in the segment string that will serve to define independent first and second syllables." [p. 33]

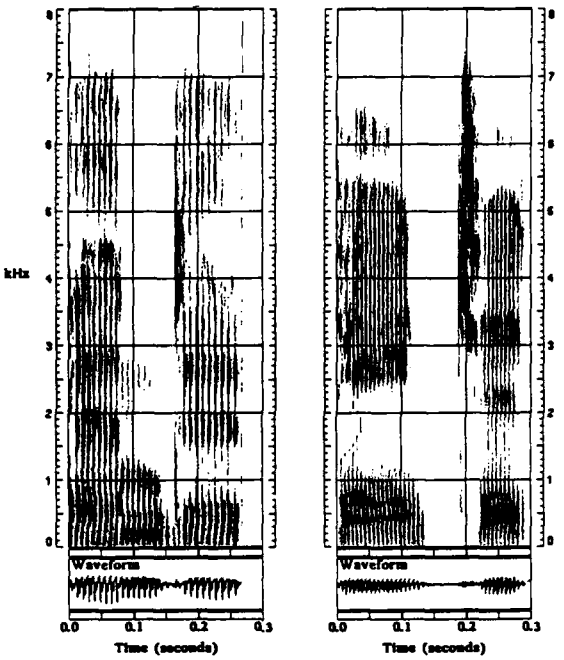


Figure 2.6: Spectrograms of /ɛndɪ/ and /entɨ/.

evenly across all frequencies. The VOT of the initial /p/ is almost 80 ms and the release is followed by a period of aspiration. In contrast, voicing begins shortly after the /b/ release. The second stop in each word occurs in a falling stress environment. The VOT of the second /p/ in "poppy" is about the same as the VOT of both of the /b/'s in "bobby." Some cues to voicing are the duration of the preceding vowel (the /ɑ/ in "poppy" is about two-thirds as long as the /ɑ/ in "bobby") and the strong voicing in the closure interval of the /b/. Place of articulation may be easier to determine than voicing for the non-initial stops.

Although studies (House and Fairbanks, 1953; Peterson and Lehiste, 1960; Klatt, 1976; Hogan and Rozsypzl, 1980) have shown that vowels are longer when they precede a voiced consonant than when they precede a voiceless one, it is unclear how useful this is for identifying stops in as limited a context as required in the previous task. Phillips (1987) had listeners label vowels presented with only the immediate phonetic context and found inter-listener agreement to be roughly 70%. Many of the errors were differences in vowel color or in the tense/lax distinction. Identification of stops in task 5, consisting

of non-syllable-initial homorganic nasal-stop sequences, may be easier than identification of singleton non-initial stops, as the nasal may encode the voicing contrast in a more accessible manner (Raphael et al., 1975). It has been observed that nasal murmurs are shorter preceding voiceless stops than voiced stops (for example, Glass, 1983; Zue and Sia, 1984). Figure 2.6 illustrates the difference in nasal murmur duration preceding a voiced and a voiceless stop. Improved identification accuracy in task 5 relative to task 4 would lend support to this hypothesis.

2.3 Database and token selection

This section describes the selection of tokens used in the listening experiments. The tokens used in the spectrogram reading experiments were a subset of the listening tokens.

The speech tokens were selected from two speech databases developed at MIT. The first is a collection of 1000 sentences recorded at MIT, referred to as the Ice Cream Database (IC). These sentences are the first 50 sets of the Harvard Lists of phonetically balanced sentences (Egan, 1944), with each set of 10 sentences spoken by one male and one female. The second corpus is a subset of the TIMIT database (Fisher et al., 1986; Lamel et al., 1986). The 2646 sentences consist of 7 sentences from each of 378 speakers, 114 female and 264 male. In the TIMIT database, each of 450 sentences was spoken by 7 different speakers. Associated and time-aligned with each sentence are an orthography, a phonemic transcription including lexical stress, word boundary, and syllable boundary markers, and a phonetic transcription. The corpora differ in the distribution of word types, style of sentences, speaker set, and recording conditions. The TIMIT database has more polysyllabic words and a wider range of sentence types than does IC. IC was recorded using a Sony omni-directional microphone, located on the chest while TIMIT was recorded using a Sennheiser close-talking microphone. Because the omni-directional microphone was able to pick up the sound radiated from tissue vibration in addition to the sound from both the oral and nasal cavities, IC has more low frequency energy for weak voiced sounds. This means that voicing during closure intervals and nasal murmurs is often stronger than in TIMIT.

The overriding concern in token selection was to have enough examples of the environments of interest, while maintaining high diversity. Since the tokens for the listening

tasks were selected from existing corpora it was not possible to balance exactly for token context within or across tasks. Thus, an attempt was made to eliminate any bias in the tokens at the cost of adding more variability. Tokens were selected by searching the phonemic transcription of the sentence to find potential regions consisting of the consonants of interest and the surrounding vowels. The phonetic and phonemic transcriptions were then compared for agreement. For example, the selection of tokens for task 1 proceeded as follows. First, all portions of the phonemic transcription matching the sequence [vowel][syllable-boundary-marker][stop][vowel] were located. Next, the corresponding regions of the phonetic transcription were checked to insure that the phonetic identity of the stop agreed with its phonemic transcription. In order to be included, each stop must have been phonetically transcribed as having both a closure interval and a release. The restriction that a stop have both a closure and a release eliminated approximately 30% of stops occurring in the contexts of interest.

After finding all the potential tokens, a subset was chosen for each task. These tokens were selected by hand according to the following "selection guidelines," aided by algorithms to assess the properties of the set. Since the recording conditions and sentence corpora are different for the two databases, an attempt was made to have equal proportions from each. Another aim was to have roughly the same number of tokens from male and female speakers and to use tokens from as many speakers as possible. Selecting tokens from as many speakers as possible helped to eliminate any speaker or sex bias. Since in both of the databases the same sentence orthography was used as the basis for the utterances of multiple speakers, an effort was made not to reuse the same portion of a sentence for different speakers. Unfortunately, for some of the rarer environments, this condition could not be met.

Table 2.1 is a summary of the token sets for each task with regard to the number of speakers, sex, and database. An attempt was made to have equal proportions of male and female speakers from each database.[4] In general there are fewer tokens from the IC database, but this is to be expected as there were less than half as many sentences as in TIMIT. Table 2.2 shows the number of distinct preceding and following vowels, and the number of distinct vowel contexts for each task. The American English vowels included were /iʸ,ɪ,eʸ,ɛ,æ,ɑ,ɔ,o,u,ʊ,ʌ,ɝ,ɑʸ,ɔʸ,ɑʷ,ə,ɨ,ɚ/. For all of the tasks, at least 15

[4]This goal was achieved for most of the tasks. However, in tasks 2 and 5, only 36% and 39% respectively of the tokens from the TIMIT database are female.

of these vowels were present in the same syllable as the stop. All 18 occurred after the stop in tasks 1 and 2, and before the stop in task 4. The total number of possible vowel contexts is 324 and the number of distinct contexts occurring for each task is shown in Table 2.2. The aim in selection was to provide enough distinct contexts for variety and coverage, while having enough samples of a given context such that the responses are statistically meaningful. If vowels are classified according to features, such as front/back or stressed/reduced, the coverage is more complete.

Table 2.1: Distribution of listening task tokens with regard to database and sex.

Task	Number of tokens	Percent TIMIT	Percent IC	Number of talkers	Percent male	Percent female
1	633	55	45	343	51	49
2	313	59	41	219	58	42
3	312	53	47	207	51	49
4	275	61	39	197	52	48
5	160	59	41	131	55	45

Table 2.2: Phonemic contexts of listening task tokens.

Task	Number of tokens	Number of preceding vowels	Number of following vowels	Number of vowel contexts
1	633	14	18	131
2	313	14	18	88
3	312	12	17	72
4	275	18	12	111
5	160	15	13	54

Chapter 3

Perceptual Experiments

In this chapter a set of perceptual experiments aimed at evaluating the listeners' ability to identify the stop consonants in a variety of local phonemic contexts are described. These experiments explored if there was sufficient acoustic information present in the extracted waveforms to allow listeners to identify the stops. Listeners were evaluated on the tasks described in Chapter 2. The remainder of the chapter proceeds as follows. In section 3.1 a summary of related work is provided. Section 3.2 describes the details of the test presentation. The perceptual results and discussion for each of the tasks individually are presented in Section 3.3, followed by cross-task comparisons.

3.1 Related work

Although over the last 40 years many experiments to evaluate the listener's perception of speech sounds have been conducted, none of the reported studies seem to address the problem of interest—the identification of the stop consonants in a limited phonemic context from multiple speakers. In this section some of the more closely related work is presented in an attempt to illustrate that point. Most of the reported studies of speech sounds in limited context were aimed at the perception of monosyllabic words or nonsense syllables (both natural and synthesized) and of speech sounds in noise. Since this research is not concerned with the effects of noise on perception, the experimental results for the best (least noisy) conditions are given. In addition, only studies using natural, not synthetic speech, are reported. Fairbanks (1958), in the well-known Rhyme Test, found listeners to achieve 89% correct identification of the initial consonant of monosyllabic words. The typical stem for the rhyming words had 8 to 9 possible completions. House

et al. (1965) reported results for a Modified Rhyme Test (MRT) where the listener chose from a closed response set of size six for consonant-vowel-consonant (CVC) words. In their experiments, listeners identified the consonant correctly about 97% of the time. Nusbaum et al. (1984) report a 96.6% identification rate for 16 consonants in CV syllables with /ɑ,i,u/. They point out that these rates are lower than found for the MRT, but that the listeners have more choices to decide among. The identification was also vowel dependent, being highest for /i/ (98.5%) and lowest for /u/ (95.2%).

As part of their investigation of the perception of speech sounds in noise and of band-limited speech, Miller and Nicely (1955) looked at perceptual confusions among English consonants preceding the vowel /ɑ/. The overall identification accuracy was 90.8% for a speech-to-noise ratio of +12 dB. The error rates within the stop class were roughly 10-15%, with the exception of /t/ which had a 2% error rate. Most of the confusions for the voiceless stops were between /k/ and /p/. The authors proposed that the high frequency burst of the /t/ separated it out, but that /p/ and /k/ needed to be distinguished by the aspiration transitions. For the voiced stops, most /b/'s were confused with voiced, weak fricatives, but /d/ and /g/ tended to be confused with each other. A later study by Clark (1983) reports an overall 95.6% open response identification rate of consonant-/ɑ/ syllables with 22 possible consonants. While identification rates for stops were not explicitly presented, stops were among the most accurately identified consonants.

Winitz et al. (1972) reported on the identification of the voiceless stops isolated from conversational speech. The stops were presented with 100 ms of the associated vowel. Correct identification was approximately 66% for the initial stops and 70% for final. This result is contrast to the finding in the earlier MRT study (House et al., 1965) that, averaged over noise conditions, stops were more accurately identified in initial position than in final position. An explanation given by the authors for the better identification of final stops is that the two talkers had experience or professional backgrounds in the speech arts and that the final stops were "perceptively released."[1] Pickett and Pollack (1964) studied at the intelligibility of words excerpted from fluent speech. They varied the speaking rate and found that the intelligibility increased with either the number of words in the sample or the sample's total duration. The word identification rate was 55% for single words, 72% for two-word sequences, and 88% for three-word sequences. The

[1]Another observation is that one of the main sources of error in initial position was /ki/—/ti/ confusions. Since no /ik/ tokens were included in the final test, no /ik/ → /it/ confusions could occur.

authors concluded that the listener needs about 0.8 sec of speech to make almost perfect identification of the words.

3.2 Experimental conditions

In this section the details of the experimental conditions for this study are presented. These include the preparation of the audio tapes used in the tests, and the test presentation.

Audio-tape preparation: The selected tokens were extracted from digitized continuous speech and randomized. Each token consisted of the stop or consonant sequence and both the preceding and following vowels in their entirety. The token was then tapered[2] and played through a Digital Sound Corporation DSC-200/240 D/A converter and recorded on a Nakamichi tape deck. Each token was played twice, with a repeat interval of one second. The inter-stimulus interval was one second for all the tasks except task 2, where it was two seconds. Listeners were allotted more time for this task because they had to make decisions about the identity of both the fricative and the stop. A block size of 10 tokens was used, with a five second inter-block interval. The first 10 tokens were repeated at the end of the tape, allowing the initial and final 5 responses to be excluded from the scoring.

Tasks 1 through 4 were divided into two parts, ranging in duration from 15 to 30 minutes long. Task 5 was about 20 minutes long. Table 3.1 shows the number of samples for each task and the approximate time for the tape. Note that the number of tokens do not add up to the total number of tokens given in Table 2.1. This is because some of the tokens (10 for task 1, and 20 for tasks 2, 3, and 4) were presented in both parts. The duplicated tokens could also be used as a check on listener consistency.

Test presentation: Ten subjects, all native speakers of American English with no known speech or hearing defects, took each experiment.[3] In this way a direct comparison

[2]The initial and final 20 ms of the waveform was tapered by half of a 40 ms Hanning window to reduce edge effects which may be distracting to listeners. Unfortunately sometimes listeners could still hear the effects of neighboring sounds.

[3]There were only nine subjects for experiment 1 because one subject did not show up on the second day.

Table 3.1: Number of tokens and tape durations for each task.

Task	Number of tokens	Tape duration (min.)
1a	340	33
1b	313	30
2a	170	17
2a	173	17
3a	170	16
3b	172	16
4a	150	15
4b	155	15
5	160	17

could be made across two tasks for a given listener. The tests were given simultaneously to five subjects in a sound treated room. The subjects listened to the test using Sennheiser HD430 headphones. The volume was adjusted to a comfortable level. At the start of each test, subjects were given an instruction sheet to read and simultaneously heard the same instructions. The instructions were followed by five example tokens, with answers supplied on the instruction sheet, and ten practice tokens, for which listeners supplied responses. No feedback was provided for the practice tokens. Subjects were given a closed set of responses to choose from and told to always provide an answer. Listeners were asked to identify the stop as one of {b d g p t k}. In task 2, listeners also identified the fricative by circling s or z on the answer sheet. In task 3, listeners chose from the set of {b d g p t k j ch}. Subjects were allowed to supply an alternate choice when they were unsure, with the hope that the alternatives supplied would provide insight into which features of the stimulus were ambiguous. Each session lasted about one hour. Since the testing was fairly tedious and long, subjects had two breaks during the session.

Different groups of subjects took each of the experiments shown in Figure 2.1. This had an added advantage of having multiple groups take tasks 1 and 4, providing more data. Experiments I, II, and III had two listening sessions each, separated by one week. The presentation order of the tests, with regard to session and task, was varied across the subject groups. Table 3.2 shows the test presentation schedule.

Table 3.2: Presentation order of the experiments to subject groups.

Experiment	Group	Number of subjects	Day 1	Day 2
I	A	4	task 1b task 2a	task 2b task 1a
	B	5	task 2b task 1a	task 1b task 2a
II	A	5	task 1a task 3a	task 3b task 1b
	B	5	task 3b task 1b	task 1a task 3a
III	A	5	task 1a task 4a	task 4b task 1b
	B	5	task 4b task 1b	task 1a task 4a
IV	A	5	task 4a,4b task 5	
	B	5	task 5 task 4b,4a	

3.3 Results and discussion

In this section the results of the perceptual experiments are presented. Although the listening tasks were organized into experiments, the results are given for the individual tasks rather than for the experiments. The subject data for each task has been pooled from the different experiments. The differences among subject groups were small compared to the variation across subjects within a group. Identification rates are compared for all the tasks, followed by an analysis of each task individually.

The listeners' ability to identify stops ranged from 85% to 97% across the tasks. The overall identification rate for each task is shown in Figure 3.1. The highest overall identification rate of 97.1%, based on 18,357 responses by 29 listeners, was obtained for the syllable-initial singleton stops of task 1. Nine subjects took the test for task 2 consisting of syllable-initial stop consonants preceded by an alveolar strong fricative. 88.3% of the 2822 responses were correct. The decline in performance from the singleton stops indicates that the presence of the fricative affected the listeners' perceptions. However,

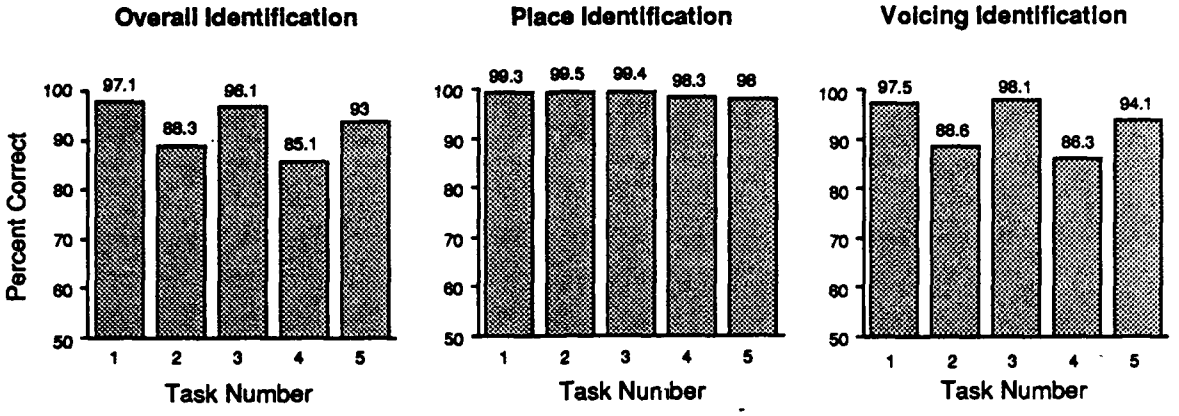


Figure 3.1: Listeners' identification rates for each task: overall, place place of articulation, and voicing characteristic.

it was not simply the presence of extra consonants that caused the performance degradation, as illustrated by tasks 3 and 5. A more detailed discussion is provided in the task-specific subsections. Syllable-initial stop-semivowel clusters and affricates were presented in task 3. The average identification rate, on 3120 responses from ten listeners, was 96.1%. As will be discussed later, most of the errors were confusions among the clusters /dr/, /tr/ and the affricates. In task 4, the syllable position of the singleton stop was varied. Twenty listeners provided 5500 responses for the non-syllable-initial singleton stops, resulting in an identification rate of 85.1%. This identification rate was substantially below that of syllable-initial singleton stops, indicating that the task was more difficult for the listeners. On task 5, where the non-syllable-initial stops were in homorganic nasal-stop clusters, the identification rate was 93% on 1488 responses by 10 subjects. There was an improvement in identification accuracy of roughly 8% from task 4 to task 5.

An analysis of the errors provides insight into what factors may be important in perception. Figure 3.2 shows a breakdown of the errors with regard to place of articulation and voicing contrast. For all tasks, the percentage of voicing errors was greater than the percentage of errors in place. In general the number of double feature errors was small—i.e., rarely were both place and voicing heard incorrectly. Listeners neglected to provide an answer less than 0.3% of the time, comprising 1% to 5% of the errors.

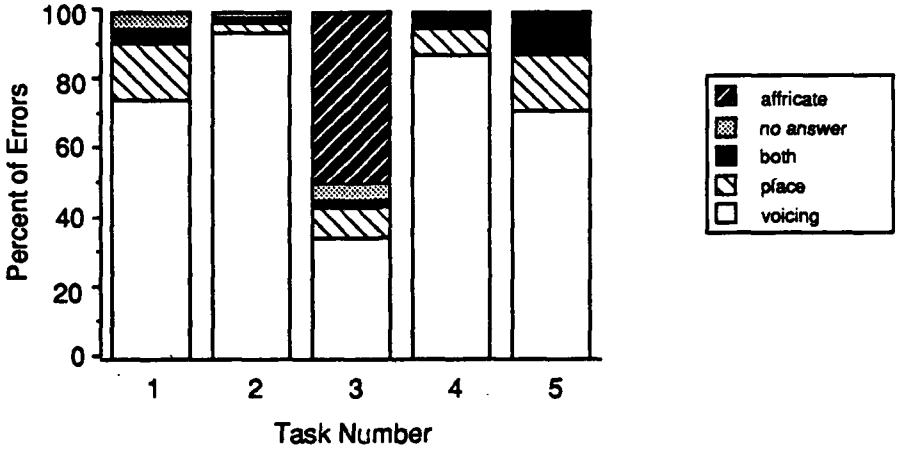


Figure 3.2: Breakdown of listeners' errors for each task according to the dimensions of place and voicing. Stop-affricate confusions are included for task 3.

With the exception of task 3, the majority of errors were in voicing only. In task 3 many of the errors were stop-affricate confusions. In fact, as shown in Figure 3.1, place of articulation was correctly identified at least 98% of the time for all tasks. The range for voicing was much larger, from a low of 86.3% for task 4 to a high of 98.1% in task 3.

In the remaining subsections, a more detailed, task-specific analysis of the errors is given. In particular, attention is focused on understanding the voicing errors, since these account for the majority of the errors. Some simple acoustic measurements provide additional insight.

3.3.1 Task 1: Perception of syllable-initial stops

A confusion matrix of the responses for task 1 is given in Table 3.3. The identification rates vary across the stops. Voiceless stops were identified more accurately than voiced stops, with better than 98% correct identification. Identification rates for the voiced stops are 96.5% for /b/ and /d/, and 94% for /g/.

Errors tended to cluster, with some tokens being misheard by several listeners.[4] Of the

[4]This clustering of errors was found for all of the tasks.

Table 3.3: Confusion matrix for listeners' identification of syllable-initial singleton stops.

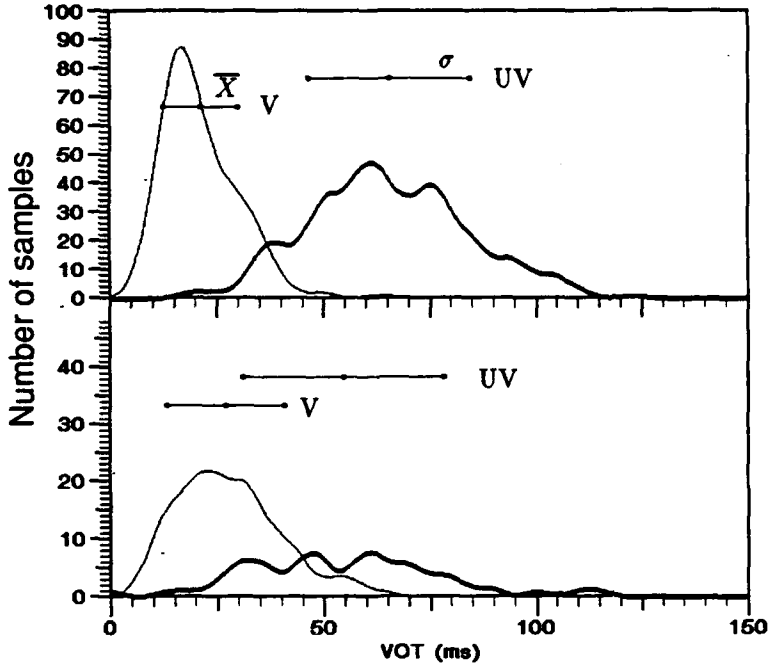
Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	110	96.5	3081	25	4	74	1		5
d	101	96.6	9	2828	23	2	59		8
g	98	94.0	7	3	2672	2		156	2
p	109	98.2	46			3103	5	1	6
t	109	98.2	1	36	9	3	3106	3	3
k	106	98.5	6		31	4	2	3028	3

633 distinct tokens 495 (78.2%) were correctly identified by all listeners. Less than half of the 138 error tokens account for more than 85% of the errors. The tokens can be divided into two sets: the first consisting of tokens that were heard correctly by all listeners (AC, for "all correct") and the second containing tokens that were misheard by at least one listener (SE, for "some error").

For this task it is proposed that the voice-onset-time (VOT) is a primary acoustic cue for voicing. While there are multiple acoustic cues that indicate the voicing feature, the supporting acoustic evidence provided here generally refers only to VOT measurements. This is in part because VOT can be measured fairly easily and reliably. In fact, the time-aligned phonetic transcriptions associated with the spoken waveforms provided the VOT directly. In addition, other acoustic cues such as the onset of the first formant, the breathiness at the onset of the vowel and presence/absence of prevoicing during closure may be correlated with the VOT (Lisker and Abramson, 1964). VOT distributions for voiced and unvoiced stops are shown in the smoothed histograms of Figure 3.3. The voicing classification was obtained from the phonetic transcription. While unfortunately this labeling is still somewhat subjective, it hopefully represents the speaker's intention.[5]

It can be seen that the distributions of VOT for voiced and unvoiced stops overlap more for SE tokens than for AC tokens. For convenience, VOT values less than 30 ms are defined as short and VOT values greater than 50 ms are defined as long. For the AC distributions, 14% of the tokens have VOT's between 30 and 50 ms. Almost 33% of the

[5]In cases where the majority of listeners were in agreement with each other but disagreed with the provided phonetic transcription, the transcription should be reconsidered. With the exception of /t/ in task 4, such disagreement between the listeners and the transcription were few (less than 3% of the tokens).



SE	N	$\overline{X}$(ms)	σ(ms)
V	220	21.3	8.8
UV	272	65.5	18.9

SE	N	$\overline{X}$(ms)	σ(ms)
V	89	27.0	13.6
UV	52	54.7	23.5

Figure 3.3: Smoothed histograms of VOT for voiced and unvoiced syllable-initial singleton stops: AC (top) and SE (bottom). The mean ($\overline{X}$) and standard deviation (σ) for each distribution are shown in the figure and provided in the table. The exploratory data analysis tool *SEARCH* (Randolph, 1985; Zue et al., 1986) was used to analyze and display the acoustic measurements. All of the histograms in this thesis were created using *SEARCH*.

SE tokens fall into the same region. Thus, if short VOT values indicate voiced stops and long VOT values indicate unvoiced stops, more SE tokens fall into the region where VOT may not provide enough information to specify the voicing feature. Roughly equal numbers of voiced and unvoiced stops have VOT values that are neither short nor long. Most of these tokens are /g/ or /p/.

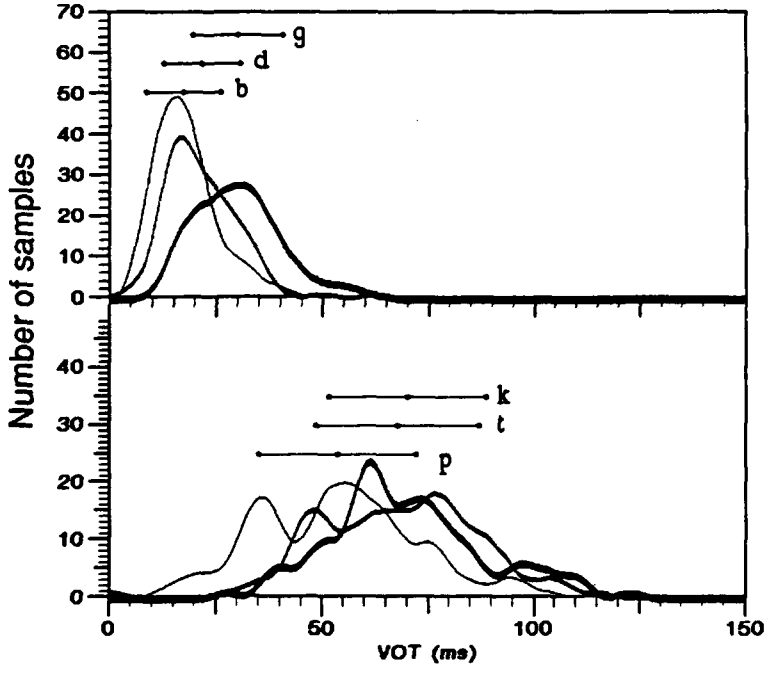
Table 3.4 shows a confusion matrix for the voicing dimension alone. The majority, almost 70%, of the voicing errors were voiced stops mistakenly called voiceless. The direction of the voicing errors was opposite of what had been expected. I expected some of the voiceless stops, particularly those preceding reduced vowels[6] to be mistaken for their voiced counterparts. However, even for stops preceding reduced vowels, the majority of

[6]Reduced vowels are those that were transcribed as a schwa, /ɨ, ə, ɚ/.

Table 3.4: Listeners' identification of voicing in syllable-initial singleton stops.

Answer	Listener's response		
	voiced	voiceless	none
voiced	8652	294	15
voiceless	129	9255	12

voicing errors were voiced stops heard as voiceless.[7] To further investigate the errors, I listened to and looked at spectrograms of the tokens preceding reduced vowels. It appears that in reduced environments the vowels are often shortened as much as, or more than, the stop release. Listeners appear to able to assess voicing correctly from the relative duration and the presence of aspiration. This direction of voicing errors (from voiced to voiceless) was seen across sex and database.



	N	$\overline{X}$(ms)	σ(ms)
b	110	17.4	8.6
d	101	21.8	8.9
g	98	30.3	10.4

	N	$\overline{X}$(ms)	σ(ms)
p	109	53.6	18.6
t	109	67.7	19.2
k	106	70.1	18.6

Figure 3.4: Smoothed histograms of VOT for syllable-initial, singleton voiced stops /b/,/d/,/g/, and voiceless stops /p/,/t/,/k/.

[7]Only 39 of the 633 tokens of singleton, syllable-initial stops preceded reduced vowels.

Although no factor obviously accounted for the direction of the voicing errors, the following observations may lend some insight. The first regards voicing in the closure interval. Many of the voiced tokens that were misheard lacked a significant amount of prevoicing in the closure interval. Over half of the voicing errors for the voiced stops occurred on /g/. /g/ has the longest VOT of the voiced stops, as can be seen in Figure 3.4, overlapping with the VOT's for the voiceless stops (Lisker and Abramson, 1964; Klatt, 1975; Zue, 1976). It was sometimes hard to tell whether or not there was aspiration or prevoicing present in spectrograms of /g/'s with voicing errors. It is possible that the long VOT, even without the presence of aspiration, causes the perception of /k/. Two-thirds of the voicing errors for /g/ occurred on back vowels, with 25% occurring for the vowel /ʌ/. The combination of a short vowel and the long VOT for /g/ may account for the higher error rate, if the listener attempts to perform some form of durational normalization.

3.3.2 Task 2: Perception of syllable-initial stops preceded by /s/ or /z/

A confusion matrix for listener responses on syllable-initial stops preceded by /s/ or /z/ is given in Table 3.5. Errors occurred on 37.1% of the 313 distinct tokens.[8] Identification was less accurate for voiced stops than for unvoiced stops. Averaged across place, voiced stops were correctly identified only 75.4% of the time, while voiceless stops were heard correctly 94.1% of the time. The majority of errors were in voicing, with two-thirds of the errors being voiced stops heard as unvoiced. This asymmetry in the errors occurred

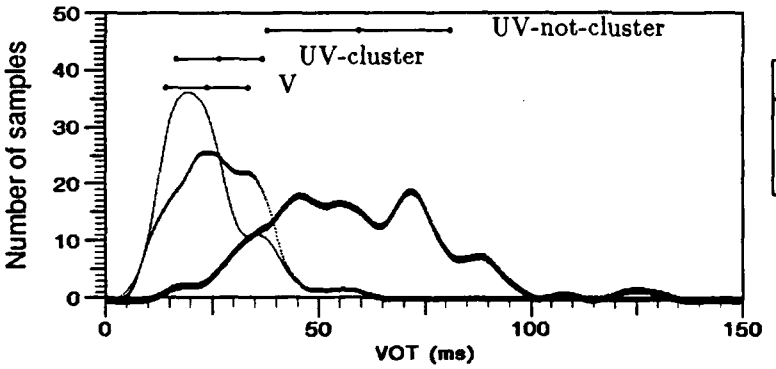
[8]There were three types of tokens in this task. When the preceding fricative is a /z/, there is a syllable boundary between the fricative and the stop. /s/ can form a syllable-initial cluster with /p,t,k/ or can occur before any of the stops if there is an intervening syllable boundary. The table below provides a breakdown of the tokens in task 2 for these three conditions. /g/ is underrepresented because it was rare in the database, as were stops following, but not in a cluster with, /s/.

/z/		/s/ not cluster		/s/ cluster	
z–b	30	s–b	13		
z–d	30	s–d	13		
z–g	7	s–g	4		
z–p	16	s–p	13	–sp	28
z–t	31	s–t	21	–st	33
z–k	32	s–k	12	–sk	30

Table 3.5: Confusion matrix for listeners' identification of syllable-initial stops preceded by /s/ or /z/.

Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	43	80.4	311	1		72	2	1	
d	43	68.7		266	4		116		1
g	11	81.8			81			18	
p	57	93.8	30			481	1		1
t	85	92.4	2	51		3	707		2
k	74	96.2			24			641	1

across sex but not across database.[9] Recall from earlier discussions (Figure 3.1) that place was correctly identified almost 99% of the time.



	N	$\overline{X}$(ms)	σ(ms)
V	97	23.7	9.7
UV-cluster	91	26.8	10.0
UV-not-cluster	125	59.3	21.5

Figure 3.5: Smoothed histograms of VOT for voiced stops preceded by /s/ or /z/, unvoiced stops in /s/-clusters, and unvoiced stops preceded, but not in a cluster with /s/ or /z/.

Several factors may affect the listeners' perception of stops preceded by /s/ or /z/. These include the role of VOT as an indicator of voicing, the voicing of the fricative, and whether or not the fricative and stop form a cluster. Figure 3.5 shows distributions of VOT for

[9]Only 39% of the voicing errors for the IC tokens were voiced stops heard as unvoiced. This is due in part to the distribution of tokens from the two databases. One of the main contributors of the voiced-to-voiceless errors were voiced stops following /s/. Only 6 of the 30 tokens of this type come from the IC database.

the three conditions: voiced stops, unvoiced stops in clusters, and unvoiced stops not in clusters. There is substantial overlap between the first two conditions, illustrating the reduced VOT of unvoiced stops in syllable-initial clusters with /s/.

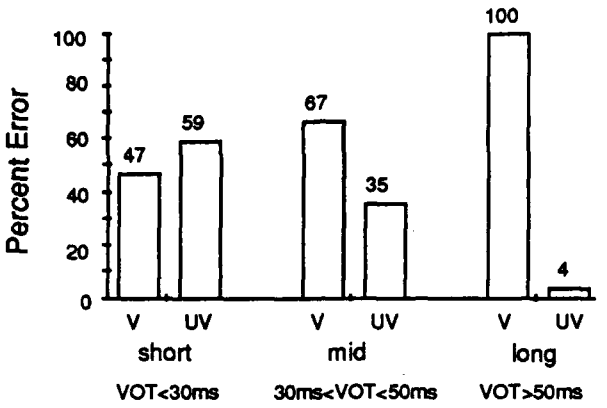


Figure 3.6: Percent of tokens misheard for short VOT (VOT < 30 ms), mid VOT (30 ms < VOT < 50 ms), and long VOT (VOT > 50 ms).

The percent of tokens on which some listener made an error is shown in Figure 3.6 for three categories of VOT: short (VOT < 30 ms), mid (30 ms < VOT < 50 ms), and long (VOT > 50 ms). If listeners used VOT as the primary cue for voicing, then most of the stops (both voiced and unvoiced) with short VOT values should have been heard as voiced. However, this was not the case. About half of the tokens with short VOT values were heard with some error for both the voiced and unvoiced stops. Listeners somehow "know" to adjust for the lack of aspiration and subsequent reduced VOT for stops in /s/-clusters. Stops with long VOT's were almost always heard as voiceless. The data for the mid-VOT stops show that more voiced stops were heard as unvoiced than unvoiced stops heard as voiced.

The ability of listeners to identify the voicing of the stop depended on the identity of preceding fricative. Figure 3.7 shows the percent of voicing errors as a function of the fricative and the location of the syllable boundary. The error rate for unvoiced stops not in a cluster with the preceding fricative was on the order of 3%, which is comparable to the error rate observed for singleton unvoiced stops in task 1. Unvoiced stops in /s/-clusters had an error rate of 10%, indicating that listeners did not always perceive the fricative and stop as a cluster. The highest error rate of 45% was observed for voiced

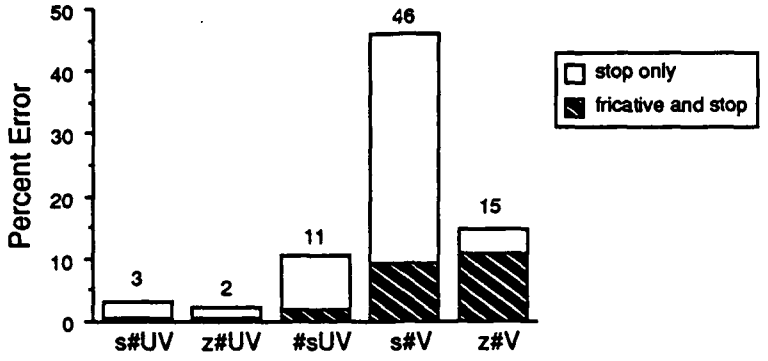


Figure 3.7: Voicing errors as a function of fricative and syllable-boundary location. Hashed regions show percent of errors where both the stop and fricative were misheard.

stops preceded by /s/. Listeners had a tendency to hear these stops as voiceless, and therefore in a cluster with the preceding /s/. When a voiced stop was preceded by a /z/, the error rate was about 15%. In 73% of these cases, the /z/ was also identified as an /s/, suggesting that the listener perceived an /s/-stop cluster.[10] The tendency for listeners to hear the voiced stops as voiceless may indicate that listeners favor syllable-internal clusters over cross-syllable boundary sequences, or may reflect a bias due to frequency of occurrence in English (Denes, 1963; Hultzen, 1965; Lamel, 1984).

Listeners appear to be able use the identity of the fricative to help identify the stop. While I was not directly concerned with the listeners' ability to identify the fricative, I had listeners make that decision in an effort to determine if the perception of voicing of the fricative influenced the voicing decision for the stop. Listeners correctly identified the fricative 78% of the time, indicating that the decision was difficult. (Only 25% of the fricatives were identified correctly by all the subjects.) The probability of perceiving the stop correctly, given that the fricative was perceived correctly, was 0.90. If the fricative was incorrectly identified, the probability of perceiving the stop correctly was 0.81.

When the listener correctly perceived the fricative as a /z/, the probability of correctly identifying the stop was 0.97, implying that listeners were able to determine that a syllable boundary occurred between the /z/ and the stop. When the stop was preceded by, but

[10]For all the other conditions, where phonotactics allow both /s/ and /z/, the fricative error rate was about 20%.

Table 3.6: Confusion matrix for listeners' identification of syllable-initial stops in clusters with semivowels and of syllable-initial affricates.

Answer	Number of tokens	Percent correct	Listener's response								
			b	d	g	p	t	k	ǰ	č	none
b	46	96.3	443	2	3	9	2				1
d	22	83.6		184			6		27	3	
g	32	99.7		1	319						
p	46	98.3	3			452	2	1			2
t	37	95.9	1				355	1		12	1
k	65	99.5			1			647			2
ǰ	32	92.2		3	1		4		295	17	
č	32	94.7					10		7	303	

not in a cluster with an /s/, the probability of perceiving the stop correctly given that the fricative was correctly identified, was only 0.77. It appears that listeners were unable to determine whether or not the /s/ and the stop formed a cluster. The probability of identifying the stop correctly, given that it was in an /s/-cluster and that the fricative was heard correctly, was 0.90. The higher probability for stops in /s/-clusters suggests there may be mutual information to help identify clusters[11] or a listener bias favoring clusters.

3.3.3 Task 3: Perception of syllable-initial stop-semivowel clusters and affricates.

This task investigated the perception of stops in syllable-initial semivowel clusters with /l,r,w/ and of syllable-initial affricates. Phonotactic constraints limit the possible syllable-initial stop-semivowel clusters. Except for /dw/ and /gw/, all of the allowable clusters were represented.[12] A confusion matrix for the listeners' responses is given in Table 3.6.

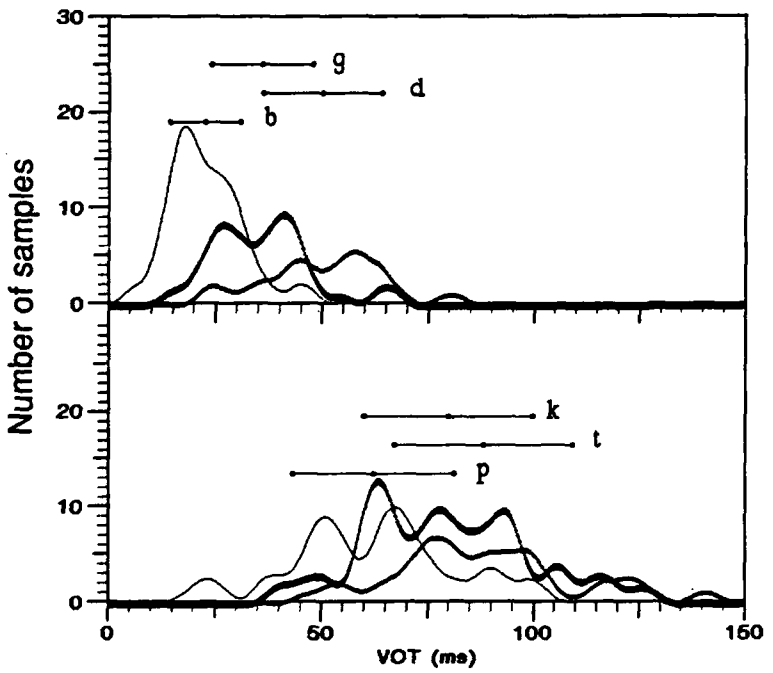
[11] It is interesting to note that the probability of hearing an /s/ correctly given that it was in a cluster was 0.89. This was higher than for /s/ not in cluster, 0.678, and for /z/, 0.771.

[12] The detailed token distribution with regard to semivowel is:

br	24	dr	22	gr	20	pr	24	tr	29	kr	24
bl	22			gl	12	pl	22			kl	25
								tw	8	kw	16
		ǰ	32					č	32		

The errors occurred on 18.6% of the 312 distinct tokens. Of the errors, 32.2% were errors in voicing only, while almost half (60) were confusions between alveolar stops and affricates. Note that only alveolar stops were ever confused with affricates.

Voiceless stops were identified more accurately than voiced. The overall identification accuracy for voiceless stops was 98.3%, while the corresponding rate for the voiced stops was 94.6%. Stops in semivowel clusters tend to have longer VOT's than singleton stops (Lisker and Abramson, 1964; Klatt, 1975; Zue, 1976). Figure 3.8 shows distributions of VOT values for the stop-semivowel clusters. The VOT's for the voiced stops are longer than those of the singleton voiced stops in Figure 3.4 by an average of 5 ms for /b/ and /g/, and 30 ms for /d/. Similar increases are seen for the voiceless stops. Since a long VOT indicates a voiceless stop, a longer VOT (as a consequence of being in a semivowel cluster) should enhance its "voicelessness." When the semivowel increases the VOT (and the amount of frication) of a voiced stop, it makes the stop seem less voiced. This may explain why almost 77% of the voicing errors were voiced stops perceived as voiceless.



	N	$\overline{X}$(ms)	σ(ms)
b	46	22.7	8.2
d	22	50.2	13.9
g	32	35.9	11.8

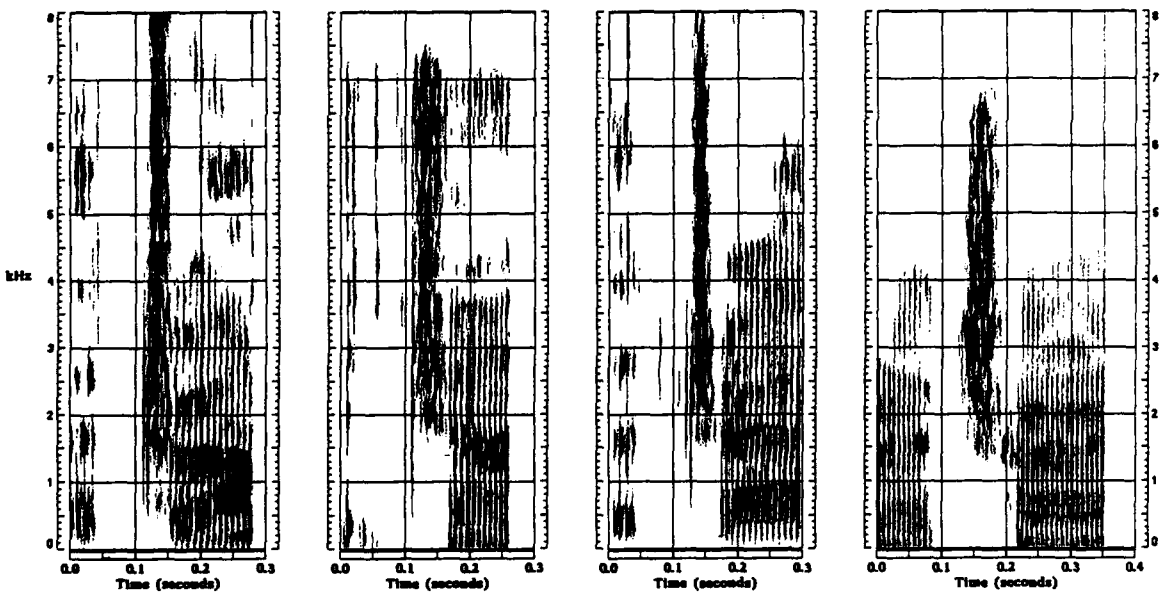
	N	$\overline{X}$(ms)	σ(ms)
p	46	62.3	19.1
t	37	88.3	21.1
k	65	79.9	19.9

Figure 3.8: Smoothed histograms of VOT for voiced stops /b/,/d/,/g/, and voiceless stops /p/,/t/,/k/ in semivowel clusters.

Table 3.7: Listeners' identification of /dr/,/tr/,/ʤ/, and /ʧ/.

Answer	Percent correct	Listener's response				
		dr	tr	ʤ	ʧ	none
dr	83.6	184	6	27	3	
tr	92.0		276		12	2
ʤ	92.2	3	4	295	17	1
ʧ	94.7		10	7	303	

The accuracy varied across the three semivowels, with the lowest error rate of 1.2% for /w/. The error rate for /l/ was 2%, with /bl/ having the most errors. Although clusters with /r/ had the highest error rate of 4.2%, this drops to less than 2% if confusions with affricates are excluded.



(a) /ədrʌ/ and /əʤʌ/ (b) /ətræ/ and /əʧɑʸ/

Figure 3.9: Spectrograms of (a) /ədrʌ/ and /əʤʌ/ and (b) /ətræ/ and /əʧɑʸ/.

Almost half of the errors were confusions between alveolar stops and affricates. Table 3.7 shows a confusion matrix for the stop-semivowel clusters /dr/ and /tr/ and the affricates. /dr/ had the lowest identification rate of 83.6%, with three-quarters of the errors being confusions with /ʤ/. Note that /ʤ/ was called /d/ only three times. Most of the /ʤ/ errors

were in voicing. All of the /tr/ confusions were with /ʧ/, while the /ʧ/ confusions were split between /t/ and /ʃ/. From these confusions, and the complete confusion matrix shown in Table 3.6, it can be seen that /dr/ was more likely to be confused with /ʃ/ than with any other stop. Similarly /tr/ was confused more with /ʧ/.

The similarity between /dr,tr/ and /ʃ,ʧ/ is illustrated in Figure 3.9. Part (a) shows spectrograms of /ədrʌ/ and /əʃʌ/. Note that the /d/ release has much more frication than is usually present in the syllable-initial singleton case. Listeners heard both of these tokens correctly. Spectrograms of /ətræ/ and /əʧɑʸ/ are shown in part (b). Five of the ten listeners confused the /ʧ/ with a /t/, while only one listener called the /t/ a /ʧ/.

3.3.4 Task 4: Perception of non-syllable-initial stops

In this task listeners identified singleton, non-syllable-initial stops. To assess the role of syllable position in stop consonant recognition, the results of this listening task are compared to the results found in task 1. The identification rate for non-syllable-initial stops was 85.1% (see Figure 3.1) as compared to 97.1% for the syllable-initial stops. A confusion matrix for the responses is given in Table 3.8.[13] The errors occurred on 52% of the distinct tokens. The errors were not evenly distributed across place of articulation. In particular, there were a striking number of voicing errors for /t/; almost 46% of the responses for /t/ were in error. The next most error-prone stop was /p/, which was misidentified almost 12% of the time.

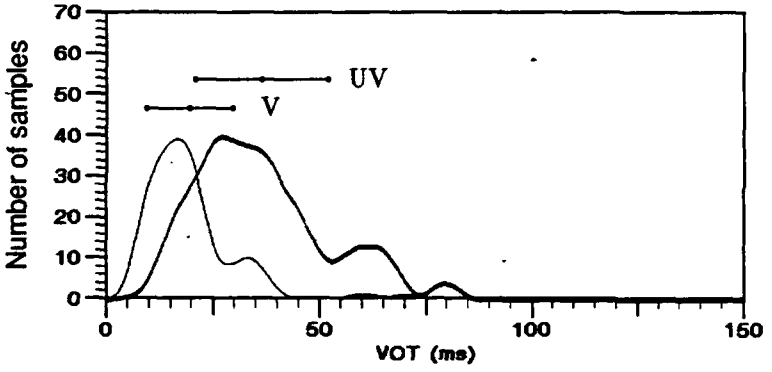
As in the syllable-initial case, most of the errors were in voicing only. In the non-syllable-initial case, almost 90% of the errors occurred on voiceless stops,[14] while in the syllable-initial case most of the errors occurred on voiced stops. This effect was seen for both databases and sexes. A possible explanation for this can be seen by looking at histograms of VOT for the voiced and the voiceless stops, as shown in Figure 3.10. The VOT's are shorter than in the syllable-initial case (see Figure 3.4), but the difference is more dramatic for the voiceless stops. The VOT's of the non-syllable-initial voiceless stops are on the order of 20 ms to 30 ms shorter than those of syllable-initial voiceless

[13] Although an attempt was made to have equal numbers of each stop, it was not possible. In particular, as seen in the number-of-tokens column of the table, /b/ and /g/ are underrepresented.

[14] Note that there were more voiceless tokens than voiced tokens in this test set. However, even normalizing for the difference in the number of tokens, over 86% of the voicing errors were on voiceless stops.

Table 3.8: Confusion matrix for listeners' identification of non-syllable-initial singleton stops.

Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	22	95.4	420	4	2	8	3	1	2
d	47	94.9	4	892	2	2	40		
g	25	95.6			478			22	
p	58	88.5	107	2	3	1027	12	8	1
t	58	54.3	4	503	13	2	630	5	3
k	65	94.8	2	1	41	7	15	1233	1

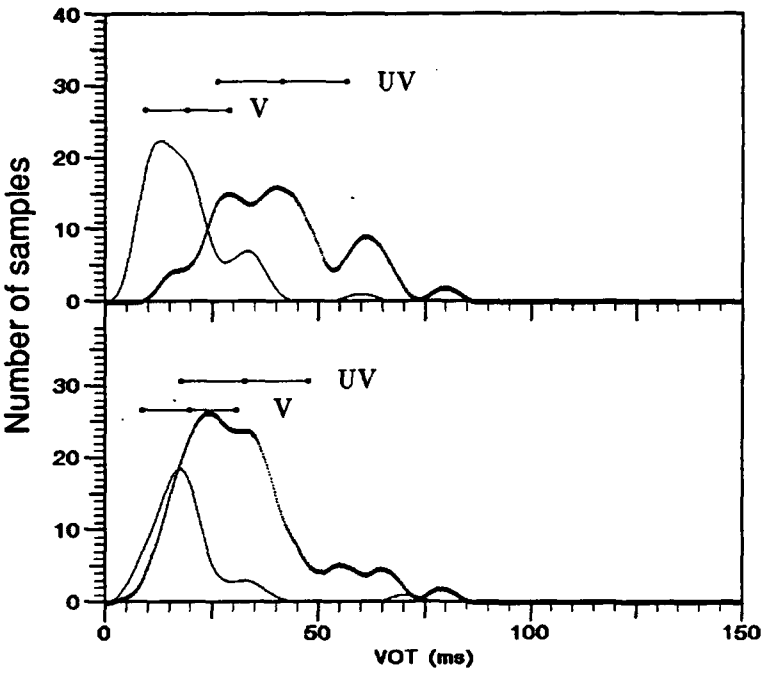


	N	$\overline{X}$(ms)	σ(ms)
V	94	19.4	10.2
UV	181	36.3	15.5

Figure 3.10: Smoothed histograms of VOT for the voiced and voiceless non-syllable-initial singleton stops.

stops. This has two consequences which may affect perception. First, there is greater overlap in the VOT distributions for non-initial stops than for initial stops. Second, many of the non-initial voiceless stops have VOT's in the range of initial voiced stops. If listeners use VOT as a primary cue for stop voicing perception, then more errors are expected for the non-initial stops.

However, stops with short VOT's were not always heard as voiced. Figure 3.11 compares the VOT's for tokens that were heard unanimously correctly to those that were heard in error by at least one listener. There is greater overlap between the distributions for voiced and voiceless stops for SE tokens than for AC tokens. In fact, the voiced stop



AC	N	$\overline{X}$(ms)	σ(ms)
V	57	19.2	9.8
UV	75	41.4	15.1

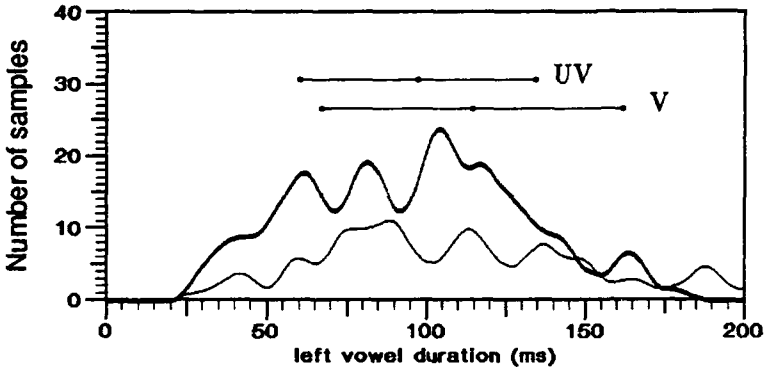
SE	N	$\overline{X}$(ms)	σ(ms)
V	37	19.8	10.9
UV	106	32.7	14.9

Figure 3.11: Smoothed histograms of VOT for the voiced and voiceless non-syllable-initial singleton stops, AC (top) and SE (bottom).

means for AC and SE tokens are the same: there is no significant difference even at the 0.2 level. On the average, for unvoiced stops, VOT's are longer for AC tokens than for SE tokens. The means are different at a significance level of 0.005.

Note however that 30% of the voiceless stops heard correctly had short VOT's. This provides evidence, along with the observation that listeners did substantially better than chance on this task, that there must be other voicing cues used by the listeners. This of course, is nothing new. Stevens and Klatt (1974) proposed that the presence or absence of a rapid spectrum change at voice-onset is an important cue for voicing of initial stops. Lisker (1957) notes that, in intervocalic position, the duration of the closure interval is shorter for voiced stops than for voiceless stops. Studies have shown that the duration of the preceding vowel is a good indicator of the voicing of a consonant (House and Fairbanks, 1953; Peterson and Lehiste, 1960; Hogan and Rozsypzl, 1980). The duration of the preceding vowel for non-initial singleton stops is shown in Figure 3.12. Since the distributions for voiced and voiceless stops overlap almost completely, it is unclear how

useful the vowel duration was for the listeners. However, the data are confounded by a variety of factors which may affect the duration and the listeners' decisions, such as stress, the tense/lax distinction, whether the vowel is a monothong or diphthong, etc. House and Fairbanks (1953) also found the fundamental frequency (F0) to be higher at the onset of voicing after a voiceless consonant. Vocal-fold vibration during the closure interval is also a cue to voicing. The relative importance of the various cues and their interactions in the perception of continuous speech is still unknown.

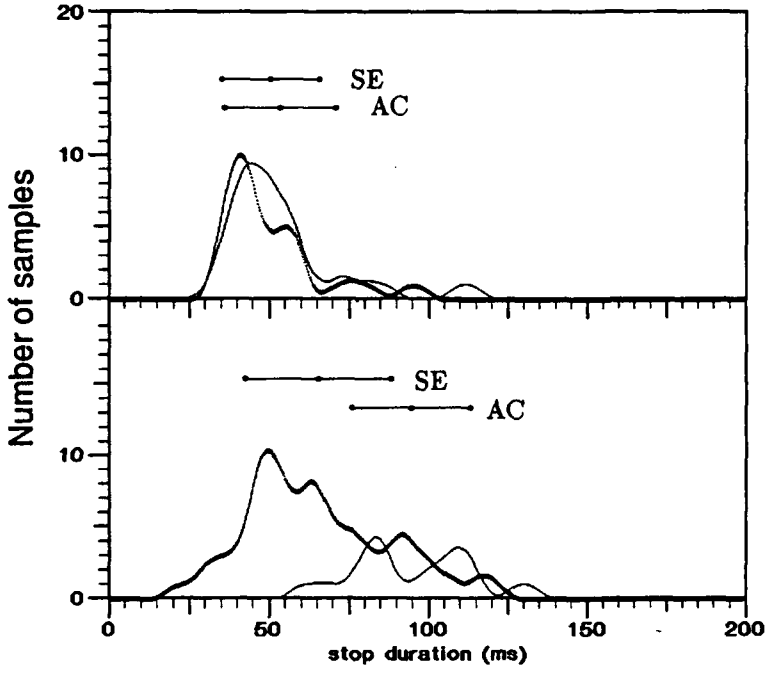


	N	$\overline{X}$(ms)	σ(ms)
V	94	114.5	47.2
UV	181	97.4	36.9

Figure 3.12: Smoothed histograms of vowel duration preceding voiced and voiceless non-syllable-initial, singleton stops.

Two types of stops were included in this task: syllable-final and ambisyllabic. The ambisyllabic stops were identified more accurately than the syllable-final stops. Ambisyllabic stops had an identification rate of 90.7% compared to 78.6% for syllable-final. Perhaps the difference is due to differences in articulation. Some of the ambisyllabic stops may be more closely associated with the following vowel and therefore have characteristics more like a syllable-initial stop. The difference in identification may also be related to stress differences: 8% of the ambisyllabic tokens follow a reduced vowel, compared to 13% of the syllable-final tokens.

As noted earlier, /t/ was correctly identified only 54% of the time. One possible explanation for the poor identification rate is that some of the /t/'s were produced as flaps. (Recall that all the tokens were transcribed as having a closure and a release, so according to the transcription, none should be flaps.) Figure 3.13 shows plots of the total stop duration for the /t/ and /d/ tokens in non-initial position for AC and SE tokens. For /d/



d	N	$\overline{X}$(ms)	σ(ms)
AC	25	53.3	17.3
SE	22	50.4	15.2

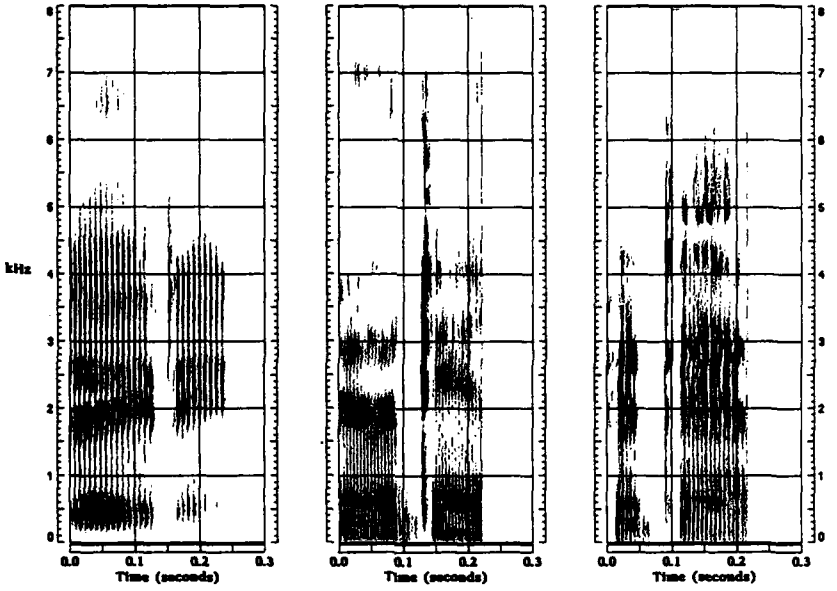
t	N	$\overline{X}$(ms)	σ(ms)
AC	14	94.6	18.6
SE	44	65.4	22.8

Figure 3.13: Smoothed histograms of total stop duration for non-syllable-initial singleton /d/ and /t/. Distribution of AC and SE tokens are shown for each stop.

there is little difference between the distributions, while a noticeable difference occurs for /t/. Many of the SE /t/ tokens overlap with /d/ in total duration. Zue and Laferriere (1979) studied the acoustic characteristics of medial /t,d/. In fact, the duration distributions for /t/ are similar to the distribution of unstressed /t/ of Zue and Laferriere. Some of the /t/ tokens may correspond to what they classified as long flaps.[15]

To check the suspicion that some of the /t/'s (or /d/'s for that matter) have turned into flaps, I listened to all of the /t/ and /d/ tokens and also looked at spectrograms of them. Particular attention was paid to those tokens that were misheard. While about 30% of the tokens sounded like flaps, only 10% of the tokens of /t/ looked like flaps in the spectrogram. An example of such a flapped /t/ is shown in Figure 3.14(a). Another 15% of the /t/'s looked a lot like /d/'s. Spectrograms of a /d/ and a /t/ that looks very similar to a /d/ are shown in (b,c). Roughly 30% of the /d/ tokens looked like flaps.

[15] According to Zue and Laferriere in long flaps "pressure has built up behind the tongue constriction and the release is accompanied by a burst of noise visible on the spectrogram." [p. 1044]



(a) (b) (c)

Figure 3.14: Spectrograms of (a) flapped /t/, (b) /d/, and (c) /t/ that looks like /d/.

3.3.5 Task 5: Perception of non-syllable-initial stops in homorganic nasal clusters

Task 5 was designed to check the hypothesis that stops in homorganic, non-initial nasal-stop clusters were easier to identify than non-initial singleton stops. The identification accuracy of 93% was an improvement of almost 8% over that of the non-initial singleton stops, supporting the hypothesis.

A confusion matrix for the responses is given in Table 3.9.[16] Errors occurred on 26.8% of the 160 distinct tokens. Of the errors, 71.4% were errors in voicing only. As in non-initial singleton stops, voiced stops were identified more accurately than voiceless.

As in task 4, both ambisyllabic and syllable-final homorganic nasal-stop sequences were included. Ambisyllabic stops were identified more accurately (95.2%) than final stops (88.9%). The difference is similar to that obtained for singleton non-initial stops, but less in magnitude.

[16] Voiced homorganic nasal-stop sequences are quite rare in the database and no tokens of /g/ satisfying the selection criteria could be found. I included extra tokens of /nd/ and /nt/, hoping that any voicing contrasts will hold for the other places of articulation.

Table 3.9: Confusion matrix for listeners' identification of non-syllable-initial stops in homorganic nasal-stop clusters.

Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	12	94.2	113	2		5			
d	57	94.7	7	540	5	1	16		1
p	26	91.9	10	7	2	239	2		
t	55	90.0		45	3	1	500	1	
k	10	96.0			4			96	

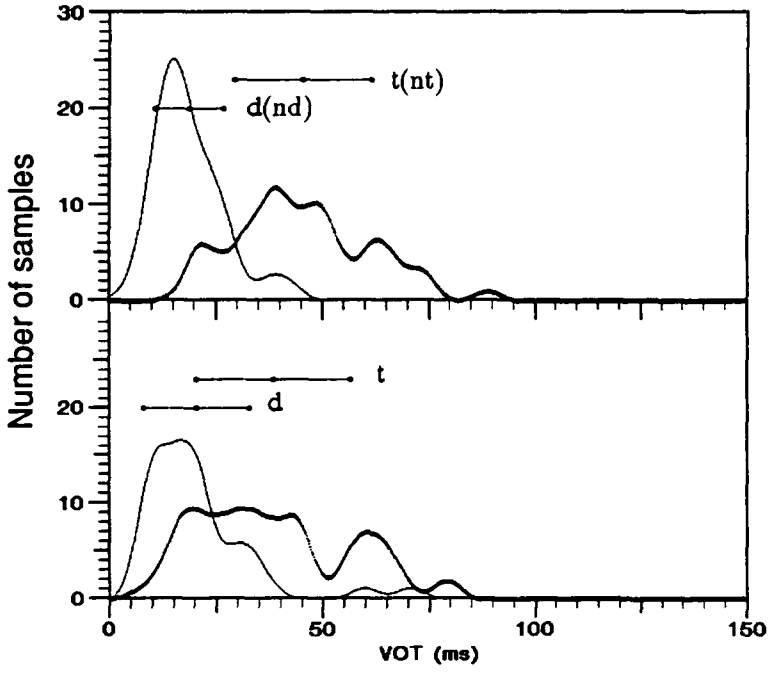
There was a notable improvement in the identification of /t/ in homorganic nasal-stop clusters relative to singleton /t/'s. While the non-initial singleton /t/'s had an error rate of almost 50%, in homorganic nasal clusters the error rate was 10%. To what can this improvement be attributed? I propose three explanations. The first is simply that the presence of the nasal in the cluster inhibits the ability to flap.[17] The second possibility is that the presence of the nasal simply increases the duration of the speech segment presented to listeners, and that the longer duration encodes more information. On the average, tokens from task 5 are 40 ms longer than tokens from task 4. According the Pickett and Pollack (1964) this difference may contribute to the improved identification accuracy. The third alternative is that the nasal is able to encode the voicing distinction in a more accessible manner than does the preceding vowel.

Figure 3.15 shows that the VOT's of non-initial /d/ and /t/ have a better separation when occurring in homorganic nasal clusters rather than as singletons. There are fewer tokens of /t/ with VOT values less than 30 ms. The difference in means for /t/ and /d/ is significant at the .005 level for both conditions. The total stop duration also has a better separation for the nasal clusters, as shown in Figure 3.16. Although the /d/'s in /nd/ clusters actually have a shorter total duration than do the singleton /d/'s, the difference is insignificant (significant only at the 0.2 level).

Information about the voicing of the stop may be encoded in the preceding nasal.[18]

[17] In some pronunciations of words like "interesting" the /nt/ turns into a nasal flap. These were excluded by the requirement that the stop transcription must have both a closure and a release.

[18] There may also be more voicing during the closure interval for the voiced stops. I did not rigorously check this possibility as it is often hard to tell where the nasal ends.



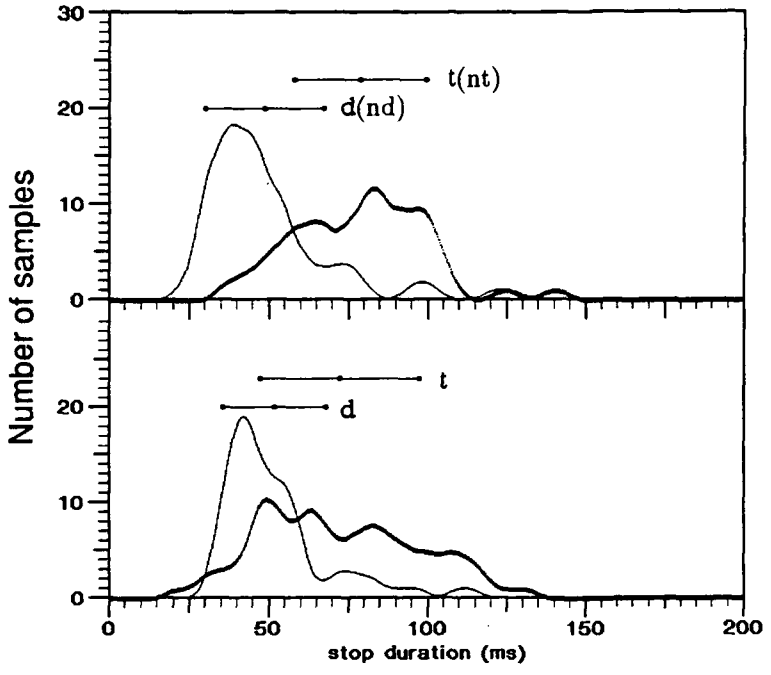
Task 5	N	$\overline{X}$(ms)	σ(ms)
d(nd)	57	18.7	7.9
t(nt)	55	45.3	15.9

Task 4	N	$\overline{X}$(ms)	σ(ms)
d	47	20.4	12.4
t	58	38.4	18.0

Figure 3.15: Smoothed histograms of VOT for /d/ and /t/ in non-initial, homorganic nasal clusters (top), and for non-initial, singleton /d/ and /t/ (bottom).

Nasals preceding voiceless stops in the same syllable tend to have a shorter duration relative to the duration of the closure interval than nasals preceding voiced stops. Figure 3.17 shows distributions of the nasal duration for voiced and voiceless stops. Although there is a significant difference in the means (α =0.005), the standard deviation of each distribution is quite large, casting doubt on the ability of listeners to use nasal duration as an indicator of voicing. A related measure, the percent nasal duration relative to the sum of the nasal and stop closure duration is shown in Figure 3.18. This measure separates the voiced and unvoiced tokens better, and thus may be a better indicator of voicing than the absolute nasal duration or the vowel duration (see Figure 3.12).

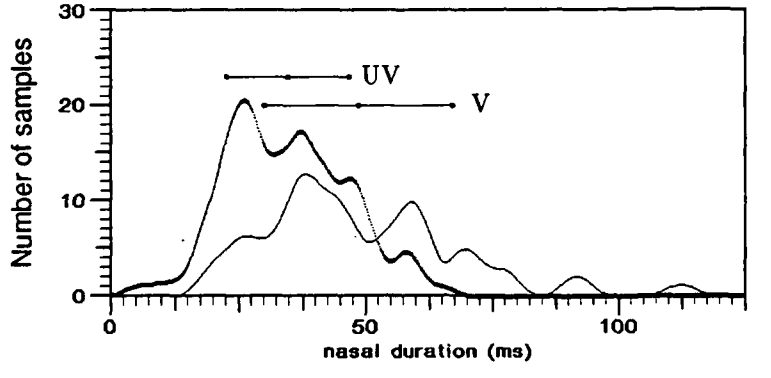
The data provide evidence for all of the proposed explanations. To further check the amount of information encoded in the nasal, I looked at spectrograms of all the tokens. In some cases the nasal duration clearly indicated the voicing of the stop. Figure 3.19 shows such an example for /nd/ and /nt/. For about 70% of the tokens, it seemed that I could judge the voicing correctly by looking at the nasal duration relative to the closure



Task 5	N	$\overline{X}$(ms)	σ(ms)
d(nd)	57	48.8	18.7
t(nt)	55	78.9	20.8

Task 4	N	$\overline{X}$(ms)	σ(ms)
d	47	51.9	16.3
t	58	72.5	25.1

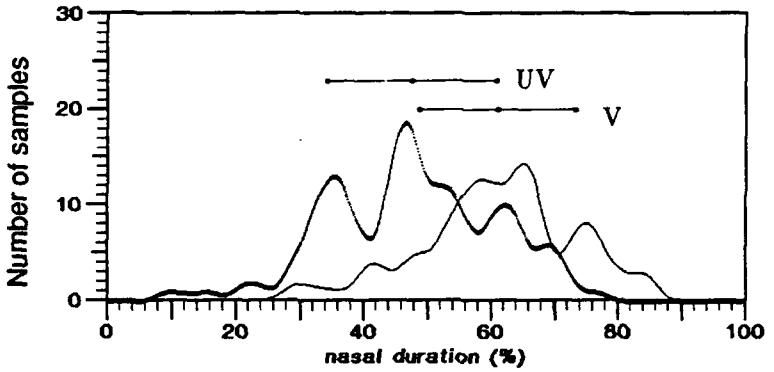
Figure 3.16: Smoothed histograms of total stop duration for /d/ and /t/ in non-initial, homorganic nasal clusters (top), and non-initial, singleton /d/ and /t/ (bottom).



	N	$\overline{X}$(ms)	σ(ms)
V	69	48.6	18.5
UV	91	34.7	12.0

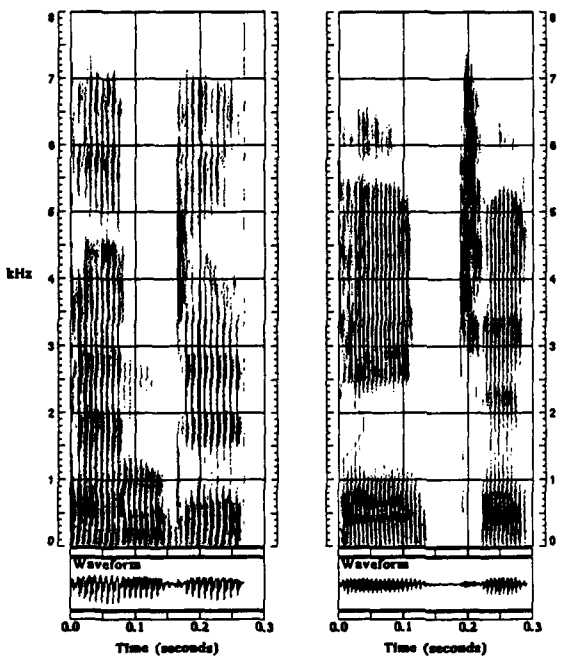
Figure 3.17: Nasal duration in voiced and voiceless non-initial homorganic stop clusters.

duration. In about 10% of the cases, the nasal information was misleading and in the remaining 20% it was ambiguous. However, I can not make any claims on what the listener is doing, only that the acoustic information seems to be available.



	N	$\overline{X}$(ms)	σ(ms)
V	69	61.1	12.2
UV	91	47.7	13.3

Figure 3.18: Relative nasal duration in voiced and voiceless non-initial homorganic stop clusters.



/ɛndɪ/ /entɪ/

Figure 3.19: Spectrograms of /ɛndɪ/ and /entɪ/.

3.4 Other factors

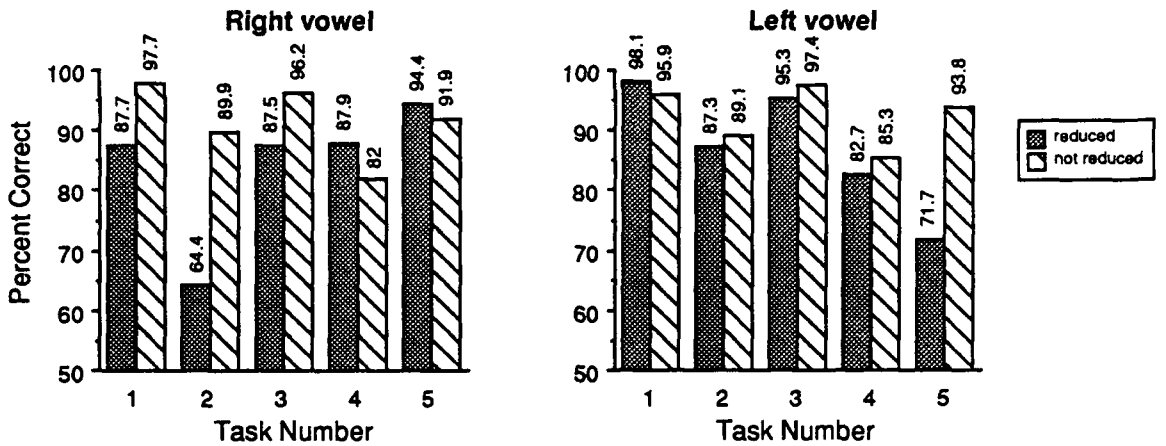


Figure 3.20: Listeners' identification accuracy of stops as a function of the stress of the right vowel and the left vowel.

Stops in reduced syllables were less well identified than those in unreduced syllables as illustrated in Figure 3.20. The identification accuracy is shown for reduced and not reduced, right and left vowels. Syllable-initial stops had higher identification accuracies when the right vowel was not reduced. The identification accuracy for syllable-initial stops was also higher when the left vowel was reduced because post-reduced stops were generally in a rising stress position. The reverse holds for the non-initial stops. They were identified more accurately following non-reduced vowels.

Figure 3.21 shows the overall stop identification rates as a function of the place of articulation and voicing characteristic of the stop. Both errors in place and voicing are included. There are some differences in accuracy across place. Velar stops had the highest identification rate for all tasks except singleton, syllable initial. As shown in Figure 3.4, it is likely that the long VOT for /g/ contributes to the high error rate. Non-initial alveolar stops had lower identification rates primarily due to confusing /t/ with /d/. This may be attributed in part to the shortened VOT and total stop duration seen for /t/ (see Figures 3.13 and 3.15(bottom).) In task 3, alveolar stops in clusters with /r/ were sometimes confused with affricates.

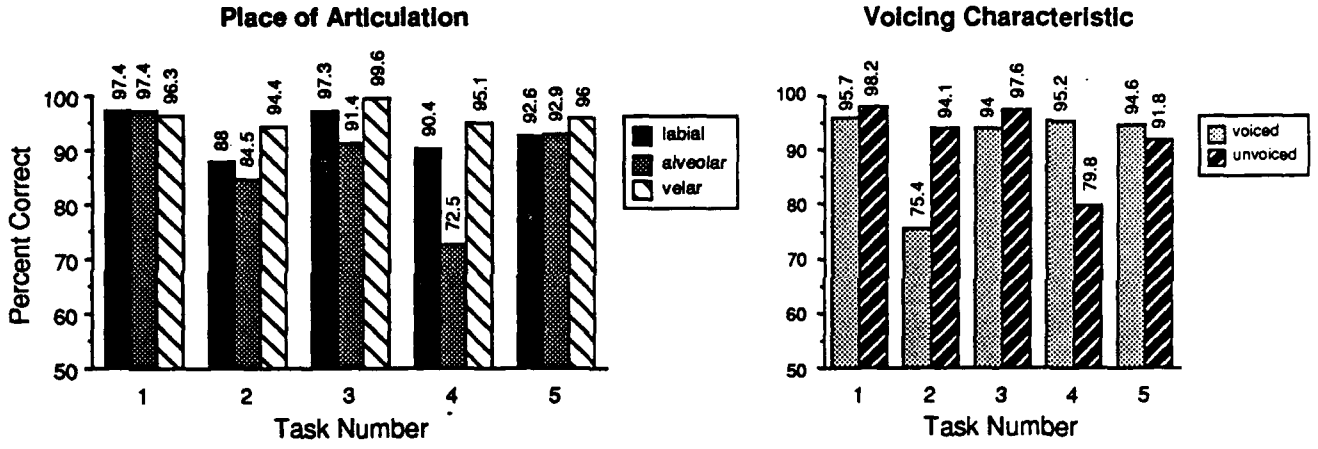


Figure 3.21: Listeners' identification accuracy of stops as a function of place of articulation and voicing.

Identification accuracy as a function of the voicing of the stop is also shown in Figure 3.21. In syllable-initial position unvoiced stops were identified better than voiced stops. For singleton stops the difference was about 2%, with many of the voiced errors occurring on /g/. When preceded by a fricative, unvoiced stops were identified almost 20% better than voiced stops. Unvoiced stops not in /s/-clusters tend to be aspirated and consequently were not as confusable. Voiced stops, on the other hand, were likely to be confused with voiceless stops unless there were strong cues to voicing in the fricative or in the closure interval. In semivowel clusters, the difference was almost 4%, and may be caused by the increase in VOT (see Figure 3.8) and frication noise. In non-initial position, voiced stops were more accurately identified than voiceless. A partial explanation may be that *non-initial voiceless stops are sometimes* unaspirated.

The effects of sex and database on the perception of stops is shown in Figure 3.22. In some of the tasks male talkers were identified better than female, however, in general, the differences were less than 1%. While tokens from IC were consistently identified better than tokens from TIMIT, the difference was only about 2%.[19] TIMIT contains

[19]The exception is task 2, syllable-initial stops preceded by the fricatives /s/ and /z/. Most errors occurred when the fricative was an /s/ not in a cluster with the stop. 80% of these tokens came from the TIMIT database, indicating that the differences observed between the databases may be token related.

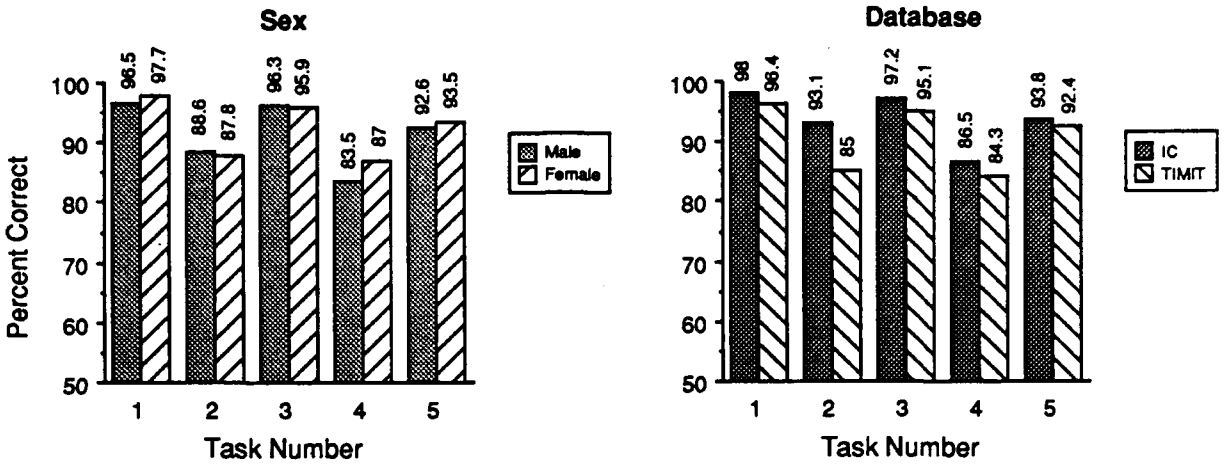


Figure 3.22: Listeners' identification accuracy of stops as a function of the sex of the speaker and the database of the token.

less carefully spoken speech than IC, with a larger variety of speakers and dialects. In addition, the words and sentences are more complicated in TIMIT than in IC (Lamel et al., 1986).

Although there were too few examples to rigorously investigate the role of vowel context on the place errors, some trends were observed. This discussion is based on tasks 1 and 4 for which there were the largest number of place errors. The most common place confusion for the singleton syllable-initial stops occurred between labials and alveolars preceding front vowels. Some of the confusions may arise from the similarity in the formant transitions into the front vowel. Since front vowels have a high second formant target, the second formant will rise from the stop into the vowel for both labials and alveolars. There were also some velar-alveolar confusions preceding front vowels which may be due to similarities in the release spectra. Other confusions were velars mistaken as labials preceding back vowels, and alveolars perceived as velars when followed by a round or retroflex vowel.

An analysis of the place errors for singleton, non-initial stops also showed some trends. For example, following front vowels, velar stops were more likely to be called alveolar than labial, while following back vowels, velar stops were called labial. Alveolar stops

were called velar more often than labial when they followed front vowels. Independent of vowel, labial stops were slightly more likely to be called alveolar than velar.

3.5 Discussion

From the data presented for tasks 1 and 4, it appears that syllable position affects stop perception. Singleton stops were identified 12% more accurately in syllable-initial position than in non-initial position. Although in both positions most of the errors were in voicing, the direction of the errors was opposite. For the initial stops, the errors were predominantly voiced stops called voiceless, whereas for the non-initial stops, voiceless stops were called voiced. The increased error rate for non-initial stops may reflect the lack of acoustic information during the release. As the non-initial voiceless stops may have reduced or no aspiration, VOT cues may be misleading. The listener may be forced to rely on other cues such as the presence or absence of voicing during the closure interval, the amount of change in the fundamental frequency, or the amount of first formant motion. The ability to identify place of articulation did not degrade in non-initial position, indicating that the release and surrounding vowels still provide sufficient information.

While the above performance differences have been attributed to syllable position, the stress environment has also changed. 66% of the syllable-initial stops occurred in a rising stress environment whereas the same percentage of the non-initial stops occurred in a falling stress environment. While differences in the stress patterns may account for some of the error differences, fewer than 10% of the tokens were in reduced syllables.

Additional consonants in clusters with a stop may affect how it is perceived. In some conditions the consonants aided perception and in others they increased the confusability. When a syllable-initial stop was preceded by an /s/ or a /z/, place of articulation perception was comparable to that of singleton stops. This implies that any place information "lost" in the preceding vowel either was compensated for by information in the fricative or was redundant. In contrast, there was almost a 10% decrease in the perception of voicing relative to singleton syllable-initial stops. In the earlier discussion data refuting the hypothesis that VOT was the most important factor for the perception of voicing were presented. However, one of the listeners, KG, appears to use the VOT

as the primary cue. KG heard stops in clusters as voiceless only 65% of the time, well below the average of 89%. KG also identified voiced stops correctly 90% of the time, as compared to 75% averaged across all listeners. The other listeners seem to be using a different strategy for perception that adjusts for "cluster." Voiceless stops not in clusters had an identification rate comparable to that for voiceless stops in task 1, indicating that a long VOT cues voicelessness for stops preceded by fricatives as well as for singleton stops.

Identification rates for the stop-semivowel clusters of task 3 were almost as high as for the singleton stops in task 1, indicating that the semivowel did not adversely affect the listeners perception of the stop. Many of the errors were due to /dr/ and /tr/ confusions with the affricates. Since affricates were not included in task 1, a direct comparison is difficult. If confusions between stops and affricates are excluded, then the error rate is about 2%. As for the singleton stops, most of the voicing errors were voiced stops perceived as voiceless. The lengthened VOT and increased frication noise for voiceless stops may increase their voicelessness, improving their perception.

In non-initial position, the presence of a homorganic nasal improved the listeners' identification of stops by almost 8% over the singleton case. The majority of the voicing errors for non-initial singleton stops occurred on voiceless stops; /t/ had a voicing identification rate close to chance. In the nasal clusters, the identification rate for /t/ was 91%. There are several factors which may contribute to the observed improvement. The presence of the nasal inhibits the tendency to turn non-initial alveolar stops into flaps. The duration of nasals in clusters with voiceless stops tends to be shorter, relative to the closure duration, than that of nasals in clusters with voiced stops. The listener may be able to use this durational information to decide voicing. In addition, the listener has more information available, both by the presence of an extra phoneme and by a longer total token duration.

In the next few paragraphs, a variety of issues related to this series of experiments are discussed.

Alternate choices: The first issue addresses the listeners' assessment of their own perception. Although listeners were encouraged to make a definitive decision, they could supply alternate choices when they were uncertain. I hoped to determine what features

Table 3.10: Listeners' responses when alternate choices were supplied.

Task	Number of multiple choices	1st choice correct (%)	2nd choice correct	comment
1	128	64	30	73% unsure of voicing
2	29	36	61	100% unsure of voicing
3	29	66	31	30% unsure of voicing, 50% alveolar–affricate
4	141	60	38	96% unsure of voicing
5	22	64	31	95% unsure of voicing

were causing the listeners' confusions by looking at their alternate choices. Table 3.10 shows the number of times a listener supplied an alternate choice, and which choice, if any was correct for each task. When multiple choices were supplied, the first choice was correct in roughly 60% of the cases. The correct answer was in the top two choices over 94% of the time. With the exception of task 3, most of the uncertainty was in voicing. Recall from Figure 3.2 that most of the listeners' errors were also in voicing. For task 3, 50% of the second choices showed alveolar stop–affricate indecision, agreeing with the largest source of errors for this task. Thus, the listeners' uncertainty, as evidenced by the second choices supplied, is in agreement with the errors they made.

Token versus response: In presenting supporting acoustic evidence, I have chosen to use the token, rather than response, as the unit of representation. (Recall, that to be AC (all correct), the stop had to be heard correctly by all listeners, whereas SE (some error), required only a single error.) This decision has the effect of weighting errors more heavily than correct responses. While accentuating the differences for "reasonable" errors, it also overemphasizes differences that may be due to listener inattention.

Task variability: Another issue regards the variability across the tasks. As mentioned previously, while selecting tokens I tried to avoid any systematic bias in the data. However, because the tokens were selected from existing databases, it was not possible to have complete coverage for all of the tasks. In particular, as the context became more explicit, fewer tokens were available. The most extensive coverage was for the singleton, syllable-initial stops of task 1, which included many vowel contexts. Since the exact

vowel environment did not seem to be an important factor in perception for task 1, the assumption was made that it was not a dominating factor in the other tasks.

Phonemic transcription: A similar problem lies in the determination of the "correct" answer. The correct answer was defined to be the phonemic transcription of the stop. Recall, that in order for a token to be included, its phonemic context had to match the search conditions and the phonetic transcription of the stop needed to agree with its phonemic transcription. A question arises as to whether or not the transcription should be considered in error, if all or many of the listeners disagreed with it. These have been counted as listener errors, but they could be considered transcription and/or production errors.

Word effects: Sometimes it was possible to hear words or parts of words in a stimulus. The perception of words arises because the speech was extracted from continuous speech using times from the utterance's time-aligned phonetic transcription. Although the transcription may place a boundary between the vowel and its neighboring phone, that phone may still influence the beginning or end of the vowel. This is particularly true in semivowel-vowel sequences where it is often difficult to place a boundary. If the listener thinks s/he hears a word, s/he may use lexical information to identify the stop. However, listeners may also think they hear words that are not there. The instructions warned the listeners of the situation: " Since these have been extracted from continuous speech, some of the things you hear may sound like parts of words or word sequences and others may not. Try not to allow your decision to be affected by what words you think you hear." I do not know how often the listeners' decisions were influenced by word perception.

3.6 Summary

These experiments were performed in order to assess human listeners' ability to perceive stops in limited phonetic environments. They also represent an effort to better understand some of the factors involved in human stop perception. As such, these experiments will serve as a baseline performance measure for comparison with spectrogram readers

and the rule-based implementation. The parallel set of human spectrogram reading experiments are discussed in the next chapter.

These perceptual studies indicate that:

- People do quite well at identifying stop consonants presented in limited phonetic context.
- Syllable position and additional consonants affect stop perception: stops were better identified in syllable-initial position. Syllable-initial errors were primarily V → UV, while non-initial errors were UV → V. Initial stops preceded by /s/ or /z/ had a higher error rate for voicing than singleton stops. For non-initial position, stops in homorganic nasal clusters were identified better than singleton stops.
- Other factors such as stress, sex and sentence corpus/recording conditions seemed to be less important. However, stops in reduced syllables were identified less accurately than those in unreduced syllables.
- There were too few errors to evaluate the role of vowel context on the place confusions.
- Errors tended to cluster on particular tokens.

Chapter 4

Spectrogram Reading Experiments

4.1 Introduction

This chapter describes a series of spectrogram reading experiments which parallel the perceptual experiments presented in Chapter 3. While previous spectrogram reading results were presented in Chapter 1, these experiments differed both in focus and extent. The aim of these experiments was to evaluate ability of spectrogram readers to identify stop consonants in a limited phonemic context, across many speakers, and in a variety of environments. Of interest were both a baseline performance measure and a comparison between spectrogram readers and listeners. Spectrogram readers were evaluated on the same tasks used to evaluate the listeners, as described in Chapter 2. For each task, the reader was presented with a set of spectrograms of tokens consisting of portions of speech extracted from sentences. As was the case for the listeners, the readers were required to identify only the stop consonant from a set of allowable choices. In all, 615 tokens, spoken by 299 speakers and extracted from over 500 sentences were identified.

Of particular interest were the following questions. Were spectrogram readers more likely to make an error when a listener made an error? Did readers make the same types of errors as the listeners? What factors or contexts affected their decisions? How did the readers interpret the spectrogram? What acoustic attributes do they use and how are they combined? In the rest of this chapter the data from the spectrogram reading experiments is interpreted in an attempt to answer the aforementioned questions. First the experimental conditions common to all the tasks are discussed. Then the spectrogram reading results and discussions for each task are presented, followed by cross-task comparisons.

4.2 Experimental conditions

Token selection: The tokens for the spectrogram reading experiments were selected from the tokens used in the corresponding listening task. The tokens for each task were divided into two subsets: one having the same proportion of tokens heard with some error as in the original task, denoted the *balanced set, (B)* and the other one heavily weighted with tokens for which listeners made errors, the *extra set, (X)*. The first subset provided a way to compare spectrogram readers and listeners on a somewhat even basis. Since listener identification was worse on the X subset, spectrogram readers were also expected to make more mistakes, providing more data on which to base an analysis of the errors.

In addition, an attempt was made to obtain the same distribution with respect to speaker sex, database, and stop identity as in the listening tests. Tables showing distributions for these factors can be found in Appendix A. Given the above constraints, the number of distinct preceding and following vowels and stress conditions were also maximized.

Spectrogram preparation and test presentation: The selected tokens were extracted from digitized continuous speech and randomized. Each token consisted of the stop or consonant sequence and both the preceding and following vowels in their entirety, as determined by the time-aligned phonetic transcription. A spectrogram was made of each token using the *Spire* facility, as described in Chapter 1. An example token of /ɪzpɛ/ is shown in Figure 4.1.

Spectrogram readers were given a training set with answers supplied so that they could familiarize themselves with the task.[1] The subjects were given a test set of approximately 50 tokens at the same time, along with written instructions. While the training and test sets given to any particular subject were non-intersecting, the test set for one subject was sometimes used as training for another. The readers were not explicitly informed about which task they were taking; however, they did receive a copy of the instructions given to the listeners from which they could deduce that information. The subjects were told to identify the consonant from the same set of choices given to the listeners. These

[1]The spectrogram readers were most familiar with reading spectrograms of isolated words or continuous sentences. Most had never participated in a test like this before. Training was self-paced; subjects were instructed to do as many of the training examples as they needed to feel comfortable with the task. Recall that the listeners also had a small amount of task familiarization. They heard 5 examples (with answers supplied on the instruction sheet) and another 10 examples for practice before each test.

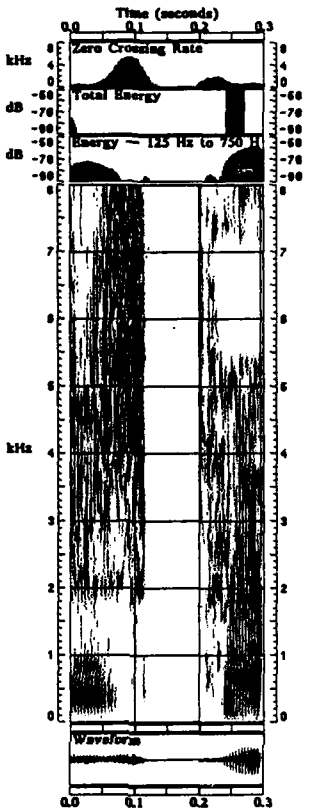


Figure 4.1: Example token of /ɪzpɛ/, as presented to spectrogram readers.

were the stops {b,d,g,p,t,k} for all tasks except task 3, which also included the affricates {ʤ,ʧ}. In task 2, readers were also asked to identify the fricative as either /s/ or /z/ and, if they could, to determine whether or not the fricative and stop formed a cluster. In the instructions subjects were encouraged to give alternate choices when they were uncertain of their decision.[2] This was natural for them, as spectrogram readers typically provide an ordered set of candidates. The alternate choices were intended to provide an indication about which features were in question.

Subjects: Five spectrogram readers participated in the experiments. The experience of the subjects varied; one subject has been reading spectrograms for about 15 years

[2]Readers were also asked to circle or to write comments about the events on the spectrogram that helped them reach their decision. The comments helped in determining which acoustic attributes were used by the spectrogram readers in forming decisions and lent some insight into how the information was combined.

Table 4.1: Number of readers and tokens for each task

Task number	Number of subjects	Number of tokens
1	5	263
2	2	102
3	1	51
4	3	153
5	1	46

investing over 3000 hours, the other four subjects have been reading spectrograms for four to eight years, estimating their experience in the range of 300 to 700 hours. All of the readers have taught the MIT version of a one-week intensive spectrogram reading course at least three times. Only spectrogram readers who had taught the spectrogram course were used as experts. While this distinction was made somewhat arbitrarily, it could be made without having explicitly evaluated the readers' performances. Table 4.1 shows the number of subjects and total number of tokens of spectrograms read for each task.

4.3 Results and discussion

This discussion parallels the discussion given for the listening experiments in section 3.3. Overall results are presented first, followed by a separate section for each task. Comparisons with the listening experiments are interdispersed throughout. While the fairest comparison between readers and listeners would be on the balanced set of tokens, all the data was used for the error analysis. The data are presented for the top-choice accuracy on all of the tokens, combining data from the B and X subsets. In section 4.5 reader performance on the two subsets is compared.

Figure 4.2 gives the overall identification rates for each of the five tasks. The bar graph represents the averaged scores for all readers. When multiple readers participated in the experiment, the best and worst accuracies are also plotted. Listener scores for the same token sets are provided for comparison. With the exception of task 2, spectrogram readers identified stops 10-15% less accurately, on the average, than did the listeners. The

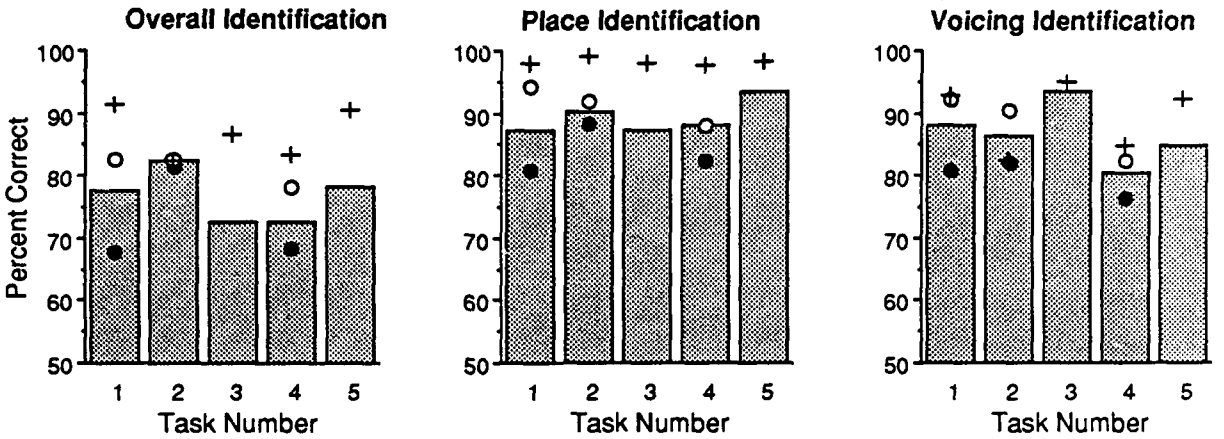


Figure 4.2: Readers' identification rates for each task: overall, place of articulation, and voicing characteristic. The bar represents the average score for all readers. The open circle (o) is the best reader's score, the filled circle (•) shows the worst reader's score, and the + denotes the average listeners' identification.

average identification rate for singleton, syllable-initial stops was 77.6%, while individual readers' rates ranged from 68-83% correct. The averaged identification rate on task 2, syllable-initial stops preceded by /s/ or /z/, was 82.3%, with a 1% difference between the readers. The one subject in task 3, consisting of stops in syllable-initial semivowel clusters and affricates, had an accuracy of 72.5%. The average correct identification for the non-initial, singleton stops in task 4 was 72.5%, with an inter-subject range from 69% to 78%. For task 5, the one subject had an accuracy of 78.3% on non-syllable-initial stops in homorganic nasal-stop clusters.

Figure 4.2 also shows the identification of the place of articulation and the voicing characteristic for all five tasks. Spectrogram readers identified place of articulation 5-10% less accurately than listeners (see Figure 3.1). The variation across tasks was also larger than for listeners: the accuracy for readers ranged from 87.1% for tasks 1 and 3 to 93.5% for task 5. Readers' identification of voicing ranged from a low of 80.4% for task 4 to 93.3% for task 3.

Another way of viewing the types of errors made by spectrogram readers is by the breakdown of the errors, as shown in Figure 4.3. In contrast to the overwhelming percentage of

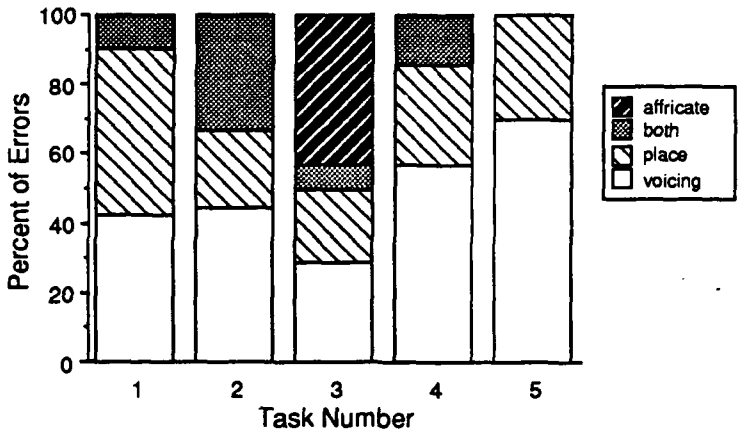


Figure 4.3: Breakdown of readers' errors for each task with regard to the dimensions of place and voicing. Stop-affricate confusions are included for task 3.

voicing errors in the listening tests (see Figure 3.2), the readers' errors were more evenly divided between place and voicing. For all tasks, a larger proportion of place errors was made by readers than was made by listeners. With the exception of task 2, the number of double-feature errors was small, i.e., rarely were both place and voicing misidentified. Almost half of the errors for task 3 were confusions with affricates.

In the remaining subsections, a more detailed, task-specific analysis of the errors is given.

4.3.1 Task 1: Spectrogram readers' identification of syllable-initial stops

A confusion matrix of the spectrogram readers' responses for task 1 is given in Table 4.2. The identification accuracy varied across the stops, with /b/ having the highest identification rate, 88.5%. Most of the /b/ errors were confusions with /d/, and usually occurred on tokens where the /b/ preceded a front vowel. /g/ had the lowest identification rate of 66.7%, being confused primarily with /k/. Listeners also made the most errors on /g/, identifying the tokens as /k/. The /p/ confusions were with both /b/ (half of them preceding reduced vowels) and /k/ (preceding back vowels). The symmetric /k/-/p/ confusion also occurred preceding back vowels. All of the /k/-/t/ confusions preceded front vowels.

Table 4.2: Confusion matrix for spectrogram readers' identification of syllable-initial singleton stops.

Answer	Number of tokens	Percent correct	Reader's response b	d	g	p	t	k
b	52	88.5	46	4	1		1	
d	40	72.5	5	29	1	1	4	
g	48	66.7		2	32	1	1	12
p	39	76.9	4	1		30	1	3
t	42	83.3		3	1		35	3
k	42	76.2			2	5	3	32

4.3.2 Task 2: Spectrogram readers' identification of syllable-initial stops preceded by /s/ or /z/

Table 4.3 gives a confusion matrix of the readers' responses for task 2. As in the singleton syllable-initial stops, /b/ had the highest identification rate of 93%. With the exception of the labials, the voiceless stops were identified almost 30% better than their voiced counterparts. This agreed with the listeners, who identified voiceless stops 10-20% better than voiced (see Appendix B, Table B.2). Almost half of the errors were in voicing; 71% of the voicing errors were voiced stops labeled unvoiced. This effect is even stronger than it appears at first, as there were more voiceless tokens than voiced. In fact, although 25% of the voiced tokens were mistakenly called voiceless, only 6% of the voiceless stops were called voiced.

Table 4.3: Confusion matrix for spectrogram readers' identification of syllable-initial stops preceded by alveolar strong fricatives.

Answer	Number of tokens	Percent correct	Readers's response b	d	g	p	t	k
b	15	93.3	14			1		
d	16	56.3	1	9		2	3	1
g	8	62.5			5	1		2
p	18	88.9		1		16	1	
t	24	87.5		2	1		21	
k	21	90.5					2	19

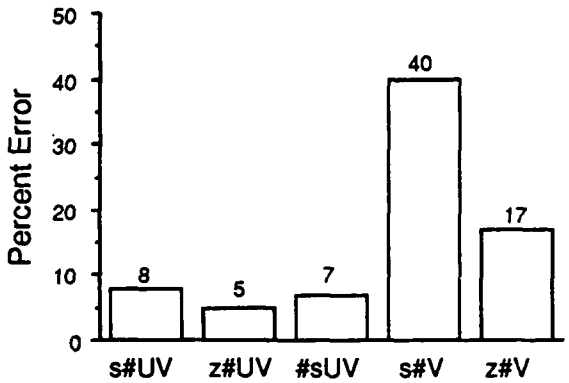


Figure 4.4: Identification of voicing as a function of the fricative and the syllable-boundary location for task 2.

The error rate as a function of the preceding fricative and the syllable-boundary location is shown in Figure 4.4. Unvoiced stops preceded by /z/ had the lowest error rate of 5%. Unvoiced stops preceded by /s/ had about the same error rate (7-8%), whether or not the stop was in a cluster with the /s/. Voiced stops preceded by /s/ had the highest error rate of 40%, which was comparable to the error rate for listeners (see Figure 3.7). In 75% of the cases where a voiced stop preceded by a /z/ was called voiceless, the fricative was also misidentified as an /s/. This suggests that the readers may have a bias towards clusters, similar to that observed for the listeners.

The readers, like the listeners, had a difficult time deciding between voiceless stops in /s/-clusters and their voiced counterparts. These account for almost 90% of the cases in which alternative choices were given. One of the readers supplied tied candidates for first choice for over half of the tokens. In 70% of these cases the alternatives were between a syllable-initial voiced stop and the corresponding syllable-initial /s/-voiceless stop cluster. The remaining 30% differed in the location of the syllable boundary or in the identity of the fricative. Not knowing how to score the ties, the reader was credited as having correctly identified the stop if either of the choices was correct. Admittedly, this boosted the score for that reader. If only half of the tied tokens were scored as correct, the reader's identification rate drops by roughly 15% for the top candidate.

Although there was a difference in listener perception of /s/-clusters, I was unable to

determine whether or not listeners were able to determine if a stop was in a cluster. In order to address this issue, the readers were asked to specify a boundary before or after the fricative whenever they could. Readers assigned a syllable boundary location 88% of the time. They correctly placed the boundary location in 73% of their attempts. Readers were most accurate (87% correct) at placing a syllable boundary before voiceless stops that were not in a cluster with the preceding fricative.

4.3.3 Task 3: Spectrogram reader's identification of syllable-initial stop-semivowel clusters and affricates

Since only one subject read the spectrograms of affricates and syllable-initial stops in semivowel clusters, the confusion matrix in Table 4.4 is rather sparse. The types of errors, however, were similar to those observed for the listeners. Alveolar stops had the largest error rate (over 50%), being confused primarily with the affricates. /dr/ was labelled as /ǰ/ more frequently than it was correctly identified. The affricates were more likely to be confused with each other, than to be called alveolar stops.

Table 4.4: Confusion matrix for spectrogram reader's identification of syllable-initial stops in clusters with semivowels and of syllable-initial affricates.

Answer	Number of tokens	Percent correct	Reader's response							
			b̥	d̥	g̥	p̥	t̥	k̥	ǰ	č
b̥	7	100.0	7							
d̥	6	33.3		2	1				3	
g̥	3	66.7	1		2					
p̥	4	75.0		1		3				
t̥	9	66.7		1			6		1	1
k̥	7	85.7				1		6		
ǰ	7	71.4					1		5	1
č	8	75.0							2	6

4.3.4 Task 4: Spectrogram readers' identification of non-syllable-initial stops

Since only one subject read the spectrograms of affricates and syllable-initial stops in semivowel clusters, the confusion matrix in Table 4.4 is rather sparse. The types of

errors, however, were similar to those observed for the listeners. Alveolar stops had the largest error rate (over 50%), being confused primarily with the affricates. /dr/ was labelled as /ʤ/ more frequently than it was correctly identified. The affricates were more likely to be confused with each other, than to be called alveolar stops.

Table 4.5: Confusion matrix for spectrogram readers' identification of non-syllable-initial singleton stops.

Answer	Number of tokens	Percent correct	Readers' response					
			b	d	g	p	t	k
b	12	66.7	8		1	3		
d	25	68.0	1	17			7	
g	11	81.8		1	9			1
p	36	72.2	5	3	2	26		
t	35	71.4		7	1	1	25	1
k	34	76.5			1	4	3	26

A confusion matrix of the readers' responses for task 4 is given in Table 4.5. Readers identified non-syllable-initial singleton stops about 5% less accurately than syllable-initial singleton stops. This difference was smaller than the 12% decrease due to syllable position observed for the listeners. Why is this so? The listeners had a particularly hard time identifying /t/ (as can be seen in Table B.4), a problem that readers do not seem to have. For example, even though only 23% of the tokens for /t/ were heard as all correct, readers correctly identified 71% of the /t/'s.

With the exception of /g/, readers identified voiceless stops better than voiced stops. This was in contrast to the listeners, who identified voiced stops more accurately by almost 15%. It appears that readers may simply have made more errors on both voiced and voiceless, and that on the remaining tokens there were enough voicing cues present for the readers to correctly deduce voicing. With the exception of /t/, readers identified all stops less accurately than listeners.

4.3.5 Task 5: Spectrogram reader's identification of non-syllable-initial stops in homorganic nasal clusters

Table 4.6 is a confusion matrix for the spectrogram reader's identification of non-syllable-initial stops in homorganic nasal clusters. The reader who participated in experiment

IV had an overall improvement of 8% from task 4 to task 5. This was the same amount of improvement observed for the listeners. While the reader identified voiceless stops perfectly, listeners made more errors on voiceless stops. The reader made the most errors for /d/; the accuracy for /d/ was only 47%. Most of /d/'s that were confused with /t/ had a weak (or non-existent) nasal murmur in the spectrogram. Since the same reader correctly identified 87.5% of singleton non-initial /d/'s, I suspect that he used a different strategy for this task, weighing the importance of the nasal murmur very highly. If a strong nasal murmur was not present, the reader called the stop voiceless.

Table 4.6: Confusion matrix for spectrogram reader's identification of non-syllable-initial stops in homorganic nasal-stop clusters.

Answer	Number of tokens	Percent correct	Reader's response					
			b	d	g	p	t	k
b	4	75.0	3			1		
d	17	47.1	3	8			6	
p	8	100.0				8		
t	15	100.0					15	
k	2	100.0						2

4.4 Other factors

Spectrogram readers identified singleton stops less accurately in reduced syllables than in syllables that were not reduced, as shown in Figure 4.5. Tasks 3 and 5 had insufficient tokens of reduced vowels to make the comparison and in task 2 the differences are small.

Figure 4.6 shows the stop identification accuracy of spectrogram readers as a function of the place of articulation and of the voicing characteristic of the stop. In syllable-initial position, labial stops had the highest identification rate. Labial stops may be easiest to identify, as they are typically weak and have a release that is spread across all frequencies. For the non-initial stops, velars were identified most accurately. The compactness of velars may be a strong cue to their identification. As shown in Figure 4.6, voiceless stops were always identified more accurately than voiced stops. The largest differences occurred for tasks 2 and 5.

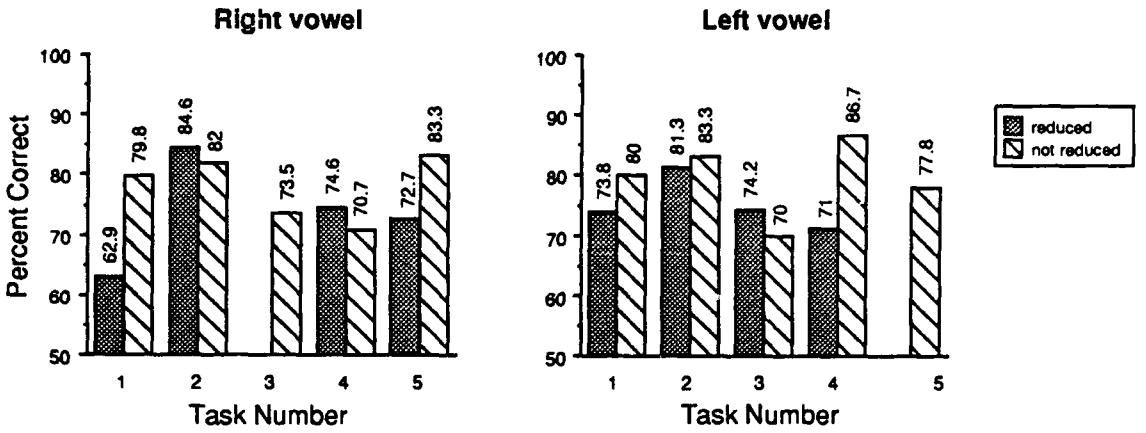


Figure 4.5: Readers' identification accuracy of stops as a function of the stress of the right vowel and the left vowel.

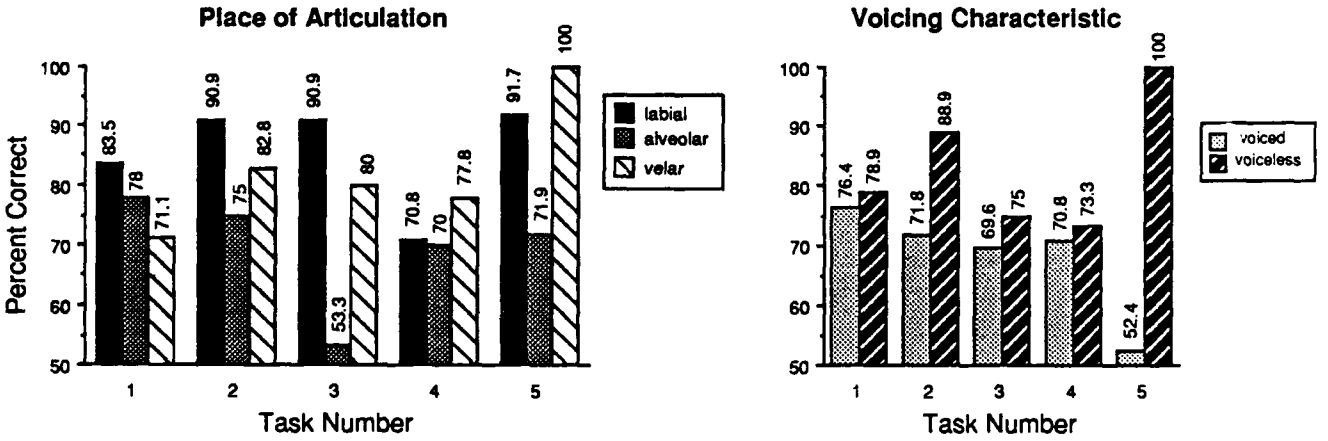


Figure 4.6: Readers' identification accuracy of stops as a function of the place of articulation and voicing characteristic of the stop.

Figure 4.7 shows the effects of sex and database on spectrogram-reader identification of stops. In all tasks, female talkers were identified better than male talkers. The difference,

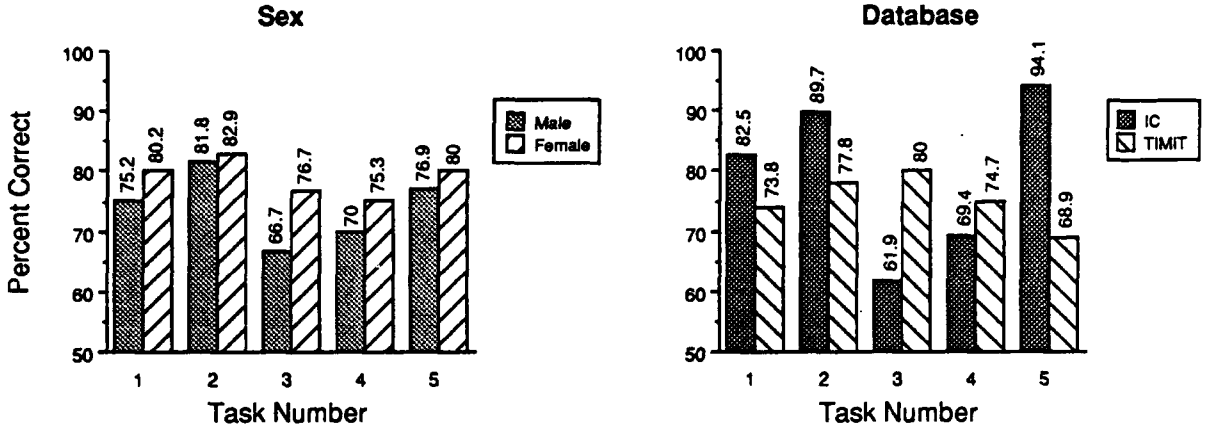


Figure 4.7: Readers' identification accuracy of stops as a function of the sex of the speaker and the database of the token.

ranging from 1% in task 2 to 10% for task 3, is larger than the 1% difference observed for listeners. Tokens from IC were identified better than tokens from TIMIT in three of the tasks.[3] Listeners consistently identified IC tokens slightly better than TIMIT tokens.

Some of the spectrogram readers' confusions on the singleton syllable-initial stops can be explained by the acoustic similarities predictable from the phonetic context. Simple acoustic-tube models can be used to illustrate the similarities due to coarticulation. These errors include: labial → alveolar preceding front vowels, where there may be high frequency energy in the release due to anticipatory coarticulation; labial-velar confusions preceding back vowels, where the energy concentration for both labials and velars is at low frequencies, near F_2 of the vowel; and alveolar-velar confusions preceding front vowels, as front velars are produced with a front cavity whose length is between that of back velars and of alveolars (Fant, 1960 [p. 187]). Although there were too few listener errors to correlate vowel context with the place errors, similar trends were observed.

[3]The differences were not simply due to the percent of tokens from each corpus that were AC. IC always had a larger percentage of AC tokens than did TIMIT. The strongest counter-example occurred for task 4, where TIMIT tokens were better identified than IC tokens. Only 27% of the TIMIT tokens were AC, whereas 47% of the IC tokens were AC.

4.5 Discussion

Previous spectrogram reading experiments: The results reported here are comparable to previously published data (Bush, Kopec, and Zue, 1983; Cole and Zue, 1980). The closest evaluation was reported by Bush, Kopec and Zue (1983). One spectrogram reader read a total of 216 word-initial stops. The stops were spoken by 6 speakers (3 male and 3 female) and occurred before the six vowels /i,eʸ,æ,ɑ,o,u/, recorded in the phrase "— is the word." The top choice accuracy was 79%, and accuracy for the top two choices was 91%. For the singleton, syllable-initial stops in this set of experiments, the same subject had a top choice accuracy of 82% and a top two accuracy of 93% (on the balanced set).[4] Thus, comparable performance by the same subject on a more varied task has been demonstrated. In fact, these experiments show that multiple spectrogram readers have somewhat comparable performance across a much larger set of speakers, phonemic environments and stress conditions. However, the conditions of the experiments were not the same and the spectrograms in the present experiment may provide more information. The speech samples in Bush et al. were lowpass filtered at 5 kHz while the speech in this experiment had a bandwidth of 8 kHz. In addition, these spectrograms were augmented by zero crossing rate, low frequency energy and total energy contours.

Cole and Zue (1980) reported on the labeling of 23 utterances spoken by two male talkers. An overall top-three candidate labeling accuracy of 85% (67% top choice only) agreement of the spectrogram reader with any one of three phoneticians was obtained. The inter-transcriber agreement was also found to be 85%. Zue's accuracy (top three candidates) on stops was 90% in word-initial position and 77% in word-medial position (see Cole and Zue (1980), Table 1.4). It is difficult to compare the results directly, as the phonetic contexts of the stops were not specified.

Performance relative to listeners: The spectrogram reader results were correlated with the listeners' performance. As expected, readers labeled AC tokens more accurately than they labeled SE tokens. The probability of all readers correctly identifying the stop ranged from 0.76 to 0.90 when the stop was AC and from 0.68 to 0.74 when the stop was SE. Figure 4.8 shows the accuracy for readers as a function of the accuracy for

[4]The same reader had a top choice accuracy of 77% and top 2 accuracy of 91% for all the tokens. The performance on the balanced set, instead of all tokens, is used for the comparison because the listener identification accuracy of 96% is closest to the 98.6% listener accuracy in Bush et al. (1983).

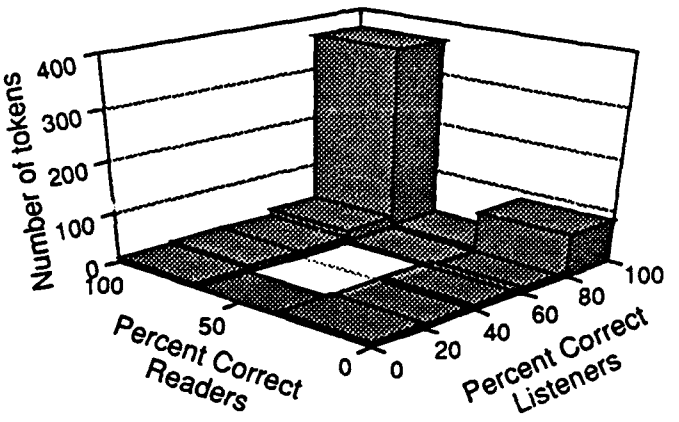


Figure 4.8: Readers' accuracy as function of listeners' accuracy.

listeners in the form of a three-dimensional histogram. There are only three values on the reader axis: 0%, 50% and 100%. This is because each token was read by at most two readers. Most tokens were read and heard correctly. The tokens that were read incorrectly often were heard incorrectly too. The errors made by the readers agreed with at least some listener in about 70% of the cases. The tendency for readers to make errors similar to those made by listeners suggests that spectrogram readers may be extracting perceptually relevant information when performing the labeling task.

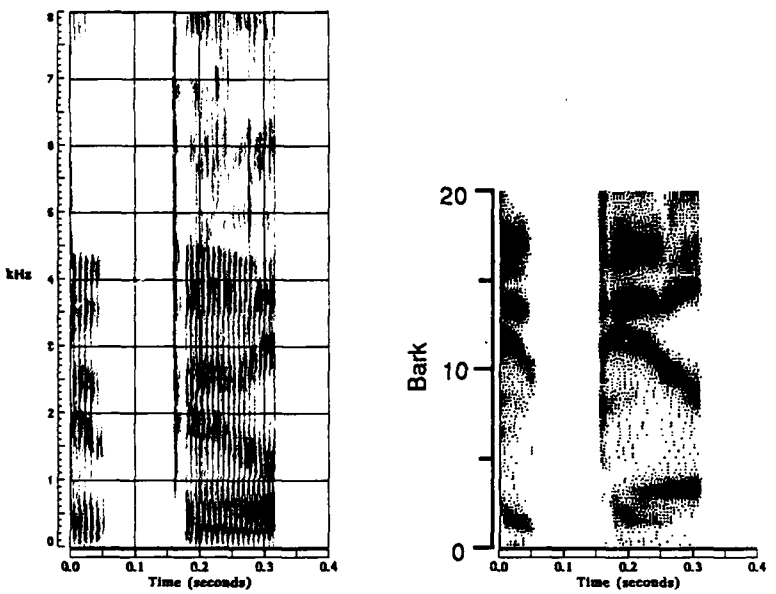
Thirty-five of the tokens were misread even though they were heard without error. These are of particular interest, as they point out where our spectrogram reading knowledge is lacking. In 22 of the cases the reader's top choice was reasonable, meaning that even knowing the answer, I might consider the readers choice best. (In all but 5 cases, I considered the reader's top choice a possible top candidate.) For the 13 voicing errors, the stop tended to have conflicting cues, such as a VOT that was long for a voiced stop, but there was no aspiration or prevoicing. (I think that 2 of the 13 voicing errors may have been oversights on the part of the reader.) The 22 place errors also occurred on tokens where there was conflicting information. Typically the burst characteristics and formant transitions appeared to be incompatible. The readers proposed the correct answer as a second choice for 19 of the tokens.

Why was the performance of the spectrogram readers consistently worse than that of the listeners? There are several possibilities including the spectrographic representation, our inability to identify and locate acoustic attributes in the spectrogram, and our inability to deduce the phonemes from the attributes. The relative importance of these factors is difficult to assess.

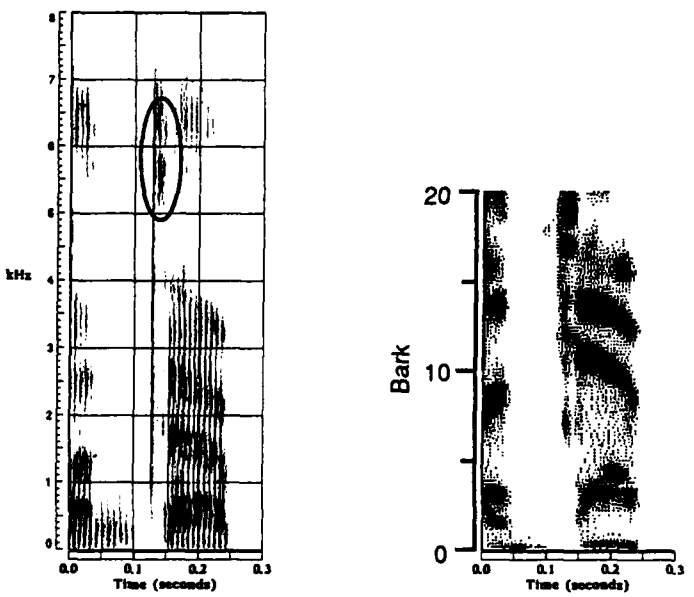
Readers were worse at identifying place of articulation than were listeners. Listeners always had 98% or better place identification. This difference may be partially due to differences in the way in which the speech signal is processed; the spectrographic analysis does not model the processing of the human auditory system. Although many aspects of human audition are not understood, some models incorporating properties of the auditory system have been developed (Lyon, 1984; Allen, 1985; Goldhor, 1985; Cohen, 1986; Seneff, 1988). The models incorporate some of the known properties such as critical-band filtering, half-wave rectification, adaptation, saturation, forward masking, spontaneous response and synchrony detection.

Perhaps an auditory-based representation can help the readers distinguish place of articulation. Figure 4.9 shows a conventional spectrogram and a spectrogram of the output of a model of the human auditory system (Seneff, 1988) for two tokens. The spectrograms on the left are the same wide-band spectrograms as have been used throughout. The spectrograms on the right are "synchrony" spectrograms. The time axes for the two are the same; but, the synchrony spectrogram is displayed on a Bark frequency scale. Spectrogram readers made place errors on both of the tokens. The stop in the spectrogram in part (a) of /ubɪ/ was identified as /d/ by both spectrogram readers. Two of the 29 listeners also identified it as /d/, but the remaining 27 correctly called it /b/. The readers described the burst as "kind of weak for a /d/" but thought the formants were "better for a /d/ than for a /b/." Since the readers did not know the identity of the vowel, it is possible that they thought the falling F_2 in the /ɪ/ (due to coarticulation with an /l/ and cued by the rising F_3) was an /u/, with F_2 raised due to coarticulation with the /d/. The synchrony spectrogram on the right accentuates the falling formant transitions, particularly into the stop from the right, supplying evidence of the labial articulation.

A spectrogram of /ədɛ/ is shown in part (b). For this token, the two spectrogram readers thought that "the release looked labial" but "the formant transitions looked alveolar" and decided the stop was a /b/. Twenty-eight of the 29 listeners correctly identified the



(a) /ubɪ/



(b) /ədɛ/

Figure 4.9: Wide-band and synchrony spectrograms of /ubɪ/ and /ədɛ/.

stop as /b/, with one listener calling the stop /p/. The synchrony spectrogram shown on the right enhances the weak, high frequency energy circled on the left, making the stop appear "more alveolar." The formant transitions in the synchrony spectrogram can be seen more clearly and are not falling into the stop as would be expected for a labial.

Table 4.7: Spectrogram readers' accuracy for all tokens, balanced subset, and extra subset. Top: top choice correct, Top 2: correct answer in the top two choices, L: Listeners' top choice identification of the same set of tokens.

Task Number	Percent correct								
	All tokens			Balanced set(B)			Extra tokens(X)		
	Top	Top 2	L	Top	Top 2	L	Top	Top 2	L
1	78	87	91	84	91	96	70	82	87
2	82	90	82	85	90	88	77	90	68
3	73	77	86	77	81	93	68	72	79
4	73	88	83	68	86	88	77	90	79
5	78	93	90	79	96	97	77	91	83

B versus X: The identification scores, averaged over all readers, are given in Table 4.7. Results are given for all tokens and for the B and X subsets separately. Reader accuracy was higher on the B tokens than on the X tokens. The lower accuracy for the X tokens was expected as they were heavily weighted with tokens on which listeners made errors (see Table A.1).

Alternate choices: Table 4.7 also includes scores for the top choice only and for the correct answer in the top two choices. For comparison, listener identification of the token subset is also provided. With the exception of task 2, the readers' accuracy was at least 10% lower than listeners' for all conditions. If the top two choices are considered, the readers' ability was closer to that of listeners'. In fact, for three of the five tasks, the readers' top two choice accuracy was better than the listeners' accuracy. Usually the choices differed in either place or voicing, but not both, so a correct partial feature specification could be provided by considering both alternatives. The second choices were almost evenly divided between place and voicing, with the exception of task 3. In task 3 their indecision was between affricates and alveolar stops.

Table 4.8: Readers' responses when alternative choices were supplied.

Task	Number of multiple choices	1st choice correct (%)	% correct in top 2	comment
1	99	63	86	50% unsure of voicing
2	58	78	91	91% unsure of voicing
3	10	70	80	40% unsure of voicing, 40% alveolar–affricate
4	102	80	93	64% unsure of voicing
5	20	65	100	75% unsure of voicing

Table 4.8 shows the number of cases in which readers supplied an alternate choice. Some readers provided alternate choices frequently, while others hardly ever gave them. In over 63% of these cases the top choice was correct. Except for task 3, the "top 2" accuracy was almost 90%.

The alternate choices reflect the readers' indecision. When only one choice was given, the readers' accuracy was 85%. When multiple choices were given the readers' top choice accuracy was 67.5%. That the readers' top choice accuracy was almost 20% better when only one choice was supplied than when multiple labels were provided indicates that readers often knew when they were uncertain. The improvement in accuracy obtained by including the second choices also shows that readers often knew which feature was uncertain.

Best reader results: It can be argued that the spectrogram readers have been penalized by presenting an average score, rather than the best score. While it was convenient to use all the responses for the error analysis, that may not be the fairest comparison between the readers and the listeners. As mentioned earlier, the spectrogram readers have had varying amounts of experience. One possibility is to compare the score for the best reader to the average score for the listeners. (All listeners are assumed to be "experts," with years of practice and hence their performance should be the same.) Since different readers participated in each experiment, the best reader for each task was independently chosen to be the reader with the highest top choice accuracy. In Figure 4.10 the scores for the best reader are shown relative to the averaged scores for the readers and the listeners scores on the balanced set of tokens. In tasks 1 and 4 the differences

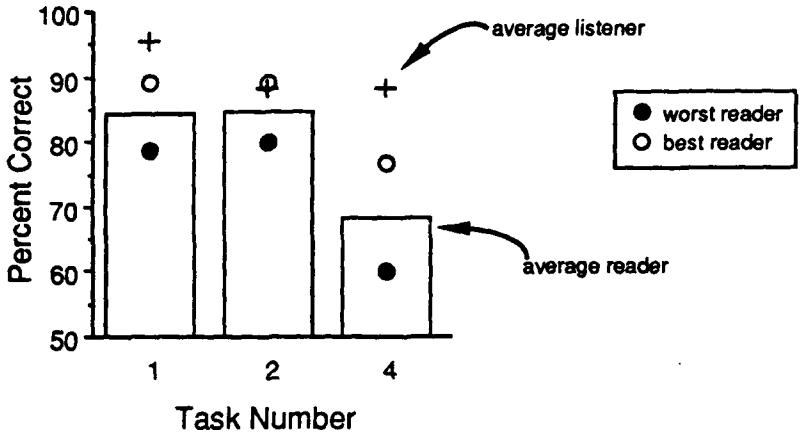
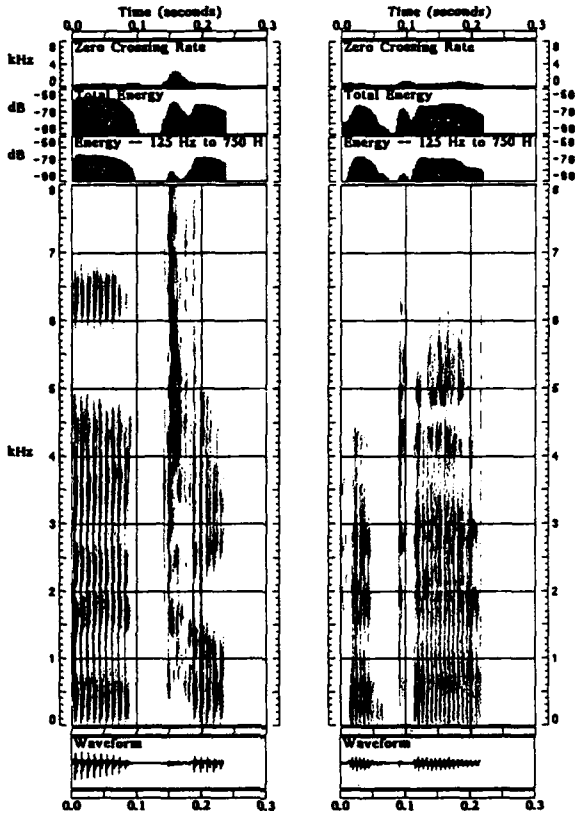


Figure 4.10: Comparison of the accuracy of the best reader and the average listener for tasks 1, 2, and 4 on the *B* tokens.

between the best reader's score and the average readers' score were on the order of 12% and 7% respectively. Thus, using the best readers' score effectively halves the difference in accuracies between listeners and readers.

Phonemic transcription: Throughout this discussion, the readers' results have been presented relative to the listeners' results. In all cases, the error rates were computed based on a comparison with the phonetic transcription of the stop. Since the phonetic transcription is subjective and may be prone to errors, some of the "errors" made by spectrogram readers and listeners may not really be errors. As an example, Figure 4.11 shows spectrograms of two tokens that were heard and read differently from how they were transcribed. The token in part (a) is a syllable-initial /d/ that was called /t/ by all 29 listeners and the one reader. The /t/ in part (b) was called /d/ by 19 of the 20 listeners and the one reader. These examples illustrate cases where the phonemic transcription of the stop may not agree with its phonetic realization.

Spectrogram readers' use of acoustic attributes: While I can not definitively conclude what spectrogram readers are doing when they read a spectrogram, I can determine some of the important acoustic attributes and, to a limited degree, infer how they

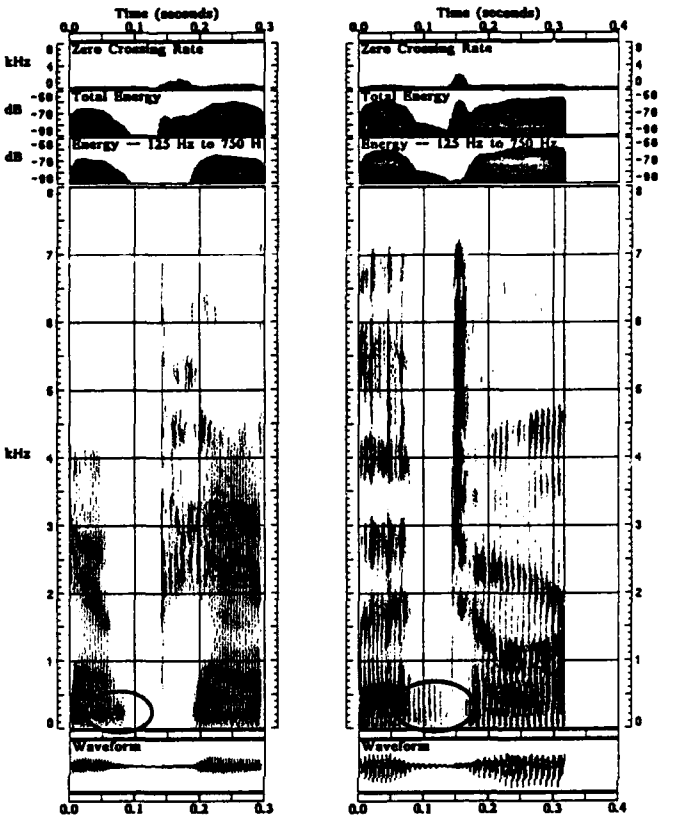


(a) /idə/ (b) /ɪtɪ/

Figure 4.11: Spectrograms of (a) /idə/ and (b) /ɪtɪ/.

are used. My inferences are based on the markings and comments made on the spectrogram by the readers, discussion of the labels with the readers, and by introspection.

Spectrogram readers tend to decide voicing and place of articulation independently. The acoustic attributes used for voicing are primarily the VOT, the presence or absence of aspiration, and the presence or absence of prevoicing during closure. For syllable-initial stops (not in /s/-clusters), VOT seems to be the most important cue. However, in noticing whether or not the VOT is short or long, readers are probably also determining whether or not the stop is aspirated, and using that information simultaneously. When the VOT is medium, readers check for aspiration, and then for prevoicing. If the stop is clearly aspirated, readers are generally willing to ignore prevoicing. When readers are uncertain about the aspiration, they weigh prevoicing more heavily— if there is prevoicing, then the stop is more likely to be voiced than voiceless. Readers write comments

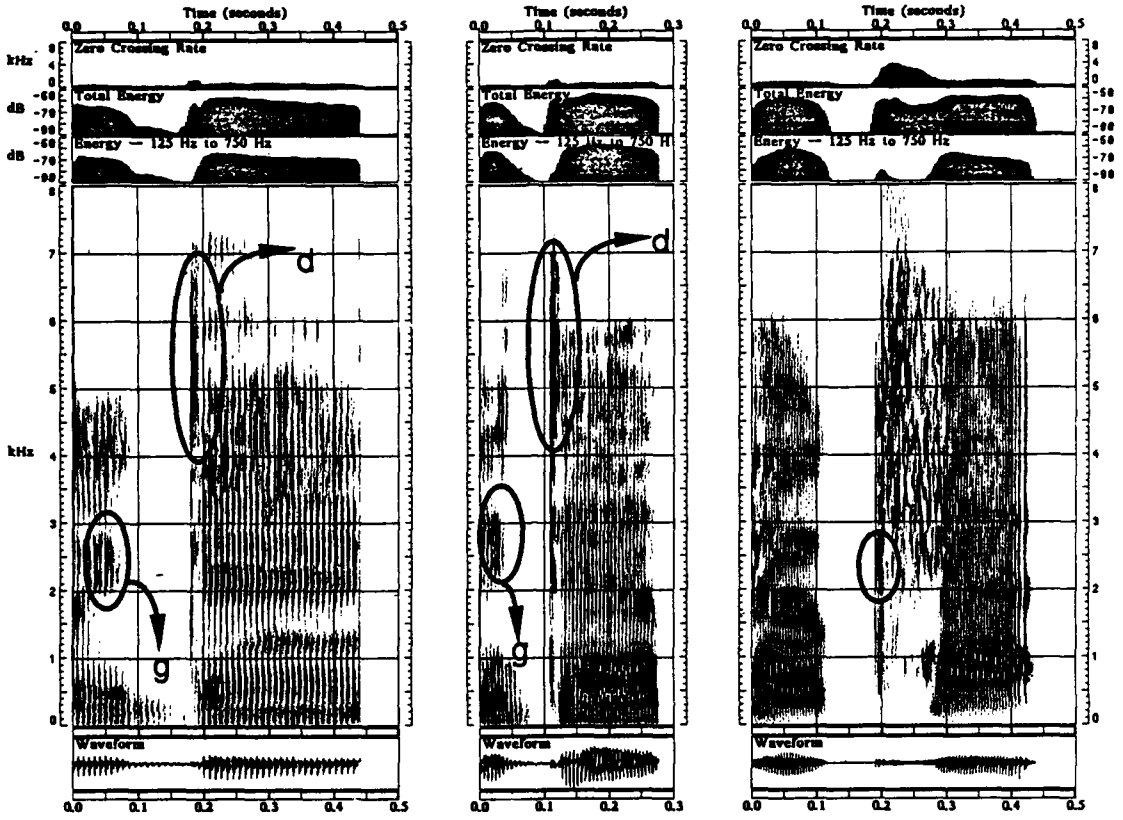


(a) /upɪ/ (b) /ədo/

Figure 4.12: Spectrograms with conflicting information for voicing.

like "VOT medium, is this aspiration? — voicing hard to tell" on their spectrograms. They also sometimes circle prevoicing to indicate they used that cue in forming their decision. Two examples with conflicting voicing information are shown in Figure 4.12. The reader circled prevoicing during the closure interval of both tokens. In part (a) the stop is clearly aspirated, and the spectrogram reader weighed that information more importantly, correctly identifying the stop as a /p/. Since this reader was particularly conservative, he also proposed a /b/ as a second choice. The stop in the spectrogram of part (b) has strong prevoicing throughout the closure. The stop has a medium VOT and it is unclear whether or not it is slightly aspirated. The strength of the prevoicing allowed the reader to correctly determine the voicing. Readers were also able to adapt for the longer VOT for stops in semivowel clusters. The readers (and the listeners) had a hard time deciding voicing for stops in clusters with /s/ and for voiced stops preceded by /s/. In non-syllable-initial position readers know that the VOT is not as reliable an indicator

of voicing. As such, they tend to pay more attention to aspiration and prevoicing.



(a) /igæ/ (b) /idæ/ (c) /okæ/

Figure 4.13: Spectrograms with conflicting place information.

From the markings made on the spectrograms by readers, I concluded that they primarily use the frequency location and distribution of the burst, the burst strength, and the formant transitions to determine the place of articulation of the stop. These three sources of information may either be confirmatory or contradictory. When they are confirmatory, such as for a labial stop that has a weak, diffuse release falling formant transitions, readers are fairly confident in their label. When the information is contradictory, readers are uncertain and tend to weigh one or two factors more heavily, disregarding the contradictory information. Three examples with conflicting information are shown in Figure 4.13. The stop in part (a) is a /g/, the second candidate given by the reader. The readers arrows indicate that he liked the release best as alveolar and the formant motion on the left as velar. In this case, the reader favored the burst location over the formants and misidentified the stop. In part (b) the same reader faced the same contradictory

information. Here the reader once again favored the burst location over the formants, and this time was correct. Knowing the correct answer, it can be argued that although the bursts for both stops look quite similar, the energy distribution for the /g/ is slightly more compact than for the /d/, and the motion of F_2 and F_3 are also better for the /g/.[5] For the spectrogram in (c), the reader saw conflicting information between labial and velar. The formant transitions on the left favor labial over velar, but the burst has a compact component, as circled, that is more like a velar. The reader decided that the formant motion and the diffuseness of the burst favored labial over velar. My suspicion is that the reader did not realize that the left vowel was an /o/ which could account for the formant transitions.

Interpreting the spectrogram readers' decisions has helped in designing the rules for the knowledge-based implementation. I discussed with each reader all of the errors that s/he made. In about 50% of the cases I agreed with the readers decision, even though I had the knowledge of what the correct answer was. In the remaining 50% I could almost always understand the error that the reader made, but knowing the answer, I could argue why the correct answer should have been considered best. In less than 2% of the cases was the error made by the reader "unreasonable."

4.6 Summary

These experiments were performed in order to assess human spectrogram readers' ability to label stops in limited phonetic environments. They also represent an effort to better understand some of the factors involved in spectrogram reading. They will serve as a performance measure in evaluating the knowledge-based system discussed in Chapter 5 and to determine which acoustic attributes are the most salient. The markings and comments provided by the readers were helpful in understanding which acoustic attributes are used, and how much weight they are given.

These spectrogram reading experiments indicate that:

[5]Since amplitude information is not particularly well captured in the spectrogram, the compactness of the release may be difficult to assess.

- Spectrogram readers were able to label stop consonants across a large number of speakers and many phonemic environments with only a limited phonetic context. The accuracy is consistent with other reported studies.

- On the average, listeners were able to identify stops 10-15% more accurately than the spectrogram readers. The difference in accuracy may be due to our incomplete knowledge of how to detect acoustic attributes in the spectrogram and how to use the attributes to form phonetic hypotheses, and in part due to inadequacies in the spectrographic representation.

- Comparing the performance of the best reader, instead of the average reader, to the average listener halves the difference in error rate.

- Syllable position and additional consonants affected the readers' ability. Singleton stops were better identified in syllable-initial position. Initial stops preceded by /s/ or /z/ had a slightly higher voicing error rate than did singleton stops. In non-initial position, stops in homorganic nasal clusters were identified better than singleton stops. Spectrogram readers confused the clusters /dr,tr/ with the affricates /ǰ,č/. These trends are the same as were observed for the listeners.

- Other factors such as stress, sex and database may be important, although the effects were hard to assess with the limited amount of data. However, singleton stops in reduced syllables were identified less accurately than those in unreduced syllables.

- Some of the place errors for the singleton, syllable-initial stops are predictable from the vowel context.

Chapter 5

Knowledge-based Implementation

Spectrogram reading entails the identification of acoustic attributes in the image, and the forming of phonetic judgements based on these attributes using our knowledge of acoustic phonetics and the articulation of speech. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. While it is impossible to know what the expert spectrogram reader is thinking as the spectrogram is interpreted, it appears that much of the knowledge can be expressed as rules. In this chapter an implementation of a rule-based system for stop identification is discussed. The implementation attempts to incorporate the knowledge used by experts in both the acoustic attributes used to describe the spectrogram and in the rules which deduce phones from these acoustic attributes. A knowledge-based system appears to be a natural medium within which to incorporate the knowledge. While heuristically-based systems for speech recognition have been developed before (Weinstein et al., 1975; Woods et al., 1976; Erman and Lesser, 1980; Espy-Wilson, 1987), using an organized framework provided by a knowledge-based system shell may make the problem more tractable. This choice provides a means of understanding how the attributes and rules interact and how the system arrives at its decisions.

The remainder of this chapter is organized as follows. First a brief discussion of knowledge-based systems is provided, followed by related work with expert systems based on spectrogram reading. Next the process of knowledge acquisition and the representation are described. A description of the qualitative acoustic attributes and the rules is given, followed by a discussion of the control strategy and an example of identifying a stop. The remaining sections discuss scoring issues and an evaluation of the system.

5.1 Background

The development of a knowledge-based system requires the existence of a lot of domain-specific knowledge and an expert that can solve the problem. While there are still many unresolved questions in the production and perception of speech, a great wealth of knowledge exists. The domain knowledge includes our understanding of articulatory principles, acoustic phonetics, and phonotactics. Past spectrogram reading experiments and those presented in the last chapter suggest that there are humans who qualify as experts. The expert must also be able to explain his/her reasoning process, as the reasoning can only be modeled indirectly on observation of the expert's own descriptions of his/her actions.

The reasoning used in spectrogram reading tends to be qualitative in nature. Qualitative reasoning is difficult to capture in statistically-based systems (Lesser et al., 1975; Jelinek, 1976; Medress, 1980). Acoustic events are either present or absent, often extend over both time and frequency, and may occur simultaneously. Researchers have argued that such acoustic events are hard to capture in systems which perform a frame-by-frame time analysis of the speech signal (Roucos and Dunham, 1987). In order to have the computer mimic the reasoning of spectrogram readers, one needs a system that can deal with qualitative measures in a meaningful way. Knowledge-based systems seem to provide this capability.

5.1.1 Knowledge-based systems

Artificial intelligence (AI), as the name implies, is an effort to develop machines or programs that exhibit what would be called "intelligent behavior" if observed in a human. Research in AI has resulted in the development of, among other applications, expert systems. The term "expert systems" (or "knowledge-based systems") describes programs that solve problems by relying on knowledge about the problem domain and on methods that human experts employ. For discussions on expert systems and knowledge-based systems see, for example, Duda and Gaschnig, 1981; Hayes-Roth et al., (1983), Buchanan and Shortliffe (1984), Waterman(1986), Grimson and Patil (1987). In solving problems, the experts use knowledge derived from basic principles and knowledge which they have acquired from experience. Perhaps the most important characteristics

of human experts are the ability to explain their conclusions, to assess the reliability of their conclusions, and to generalize their knowledge. Experts integrate multiple sources of information and often reason about qualitative data. The goal of many researchers in the expert/knowledge-based systems field is to develop programs modeled on human cognitive behavior.

A knowledge-based system explicitly separates problem solving strategies and the domain knowledge. The domain knowledge is usually expressed in a set of statements and/or principles. A typical system consists of a general inference engine and a database that is specific to a particular task. The inference engine keeps track of current hypotheses, applicable rules, and a history of rules that have already been applied. Knowledge-based systems often have multiple levels of abstraction making them easy to use and modify (Hayes-Roth, 1984). Much of the processing is symbolic rather than numeric.

In rule-based systems, a subset of knowledge-based systems, the conclusion is based on the results of a set of *if-then-else* rules, operating on some input data or initial conditions. Two control strategies are widely used: forward chaining and backward chaining. Forward chaining (or data-directed reasoning) is reasoning from facts to form conclusions. Backward chaining (or goal-directed reasoning) is reasoning from a desired conclusion backward to the required facts. Some systems integrate both mechanisms (Waterman and Hayes-Roth, 1978). Some systems attempt to reason under uncertainty, to deal with qualitative information, and to combine different sources of information in meaningful ways (Kanal and Lemmer, 1986). Some systems can deal with partial and/or conflicting information and can handle multiple hypotheses (Pauker et al., 1976). While still an active area of research, knowledge-based systems have been applied to a variety of applications including medical diagnosis (Szolovitz, 1982; Clancey and Shortliffe, 1984), business (Winston and Prendergast, 1985), mineral exploration (Duda et al., 1981) and others (Davis et al., 1977; Gaschnig, 1982; McDermott, 1982; Miller et al., 1982).

Perhaps the most important reason for using a knowledge-based system is that the knowledge is represented explicitly in the facts and the rules. The user can ask the system to explain the reasoning used to obtain its answer, providing a way to evaluate the rules and to understand interactions among them. In addition, interactively working with the system may help elucidate the reasoning used by human experts.

5.1.2 Related work

There have been several attempts to design speech recognizers that model spectrogram reading over the past five years. In this section a brief summary of these attempts is provided in chronological order. Some of the systems have been implemented and evaluated, while others were just proposed.

Johanssen et al. (1983) proposed an automatic speech recognition system based on spectrogram reading. Their proposed system, SPEX (spectrogram expert), consisted of three interacting experts: a visual reasoning expert that identifies important visual features in the spectrogram, providing a symbolic description of the utterance; an acoustic-phonetic expert that reasons about how visual features relate to phonemes; and a phonetics expert that reasons about allowable phoneme sequences and produces an English spelling from a string of phonemes.

Another project was reported by Johnson et al. (1984). A knowledge-based system, implemented in Prolog, had rules to capture the relationship between acoustic events as represented in the spectrogram and linguistic units, such as phonemes. Features from the spectrogram were manually caricatured for each "area," (where it is assumed that "area" refers to some manually defined acoustic segment) and supplied to the system. The system was tested on twenty sentences spoken by two male and two female speakers. 63% of the phonemes were correctly identified, 21% missed, and 16% confused with other phonemes. The authors reported that they expect the system performance to improve as the rules are refined and that they were encouraged by the system's high ability to discriminate natural classes.

Carbonell et al. (1986) have implemented a system, APHODEX (acoustic phonetic decoding expert), for reading French spectrograms. (See also Carbonell et al., 1984; Haton and Damestoy, 1985.) A set of pre-processors performed a coarse segmentation of the acoustic signal and classified the segments as vowels, plosives, or fricatives, The output of the preprocessors was stored in a "fact base" associated with each segment. A set of inference rules attached a label, or list of labels, to each segment and may have refined the original segmentation. The inference engine analyzed the segments in an utterance from left-to-right, using both forward and backward chaining. Only rules applicable to the phonetic class of the segment were applied. Backward chaining was used

when the identity of a segment was conditioned by the phonetic nature of its right-side context.

Knowledge was stored in the APHODEX database using *frames*. Frames (Minsky, 1975) are data structures for representing stereotyped situations in which features of an object are associated with *slots* and *slot-values*. Two types of frames were used during recognition: prototypes and instances. Prototypes associated phonetic units with a set of expected acoustic characteristics defined in advance by the system designer. Instances were segments of the utterance created during recognition. The slots of the instances were filled in by the rules. An instance was identified by comparing it to all prototypes in its class, and choosing the prototype that matched it best. The authors note that at the time of writing the knowledge base had 200 rules. However, no evaluation was reported.

As part of his doctoral research, Stern developed an expert system to identify a subset of French sounds (Stern, 1986; Stern et al., 1986). The aim was to formalize and test the knowledge used in spectrogram reading. The knowledge was formalized in a set of production rules for acoustic, phonetic, and phonotactic information. The acoustic knowledge was encoded in rules of the form:

> **If** *feature description* **and** *context*
> **then** weighted phonetic subset or phoneme.

Phonotactic information was used to suggest hypotheses and to reduce the search space. The phonetic knowledge was integrated in a phonetic network representing the phonetic relationships between classes. A forward chaining inference engine, implemented in Prolog, used a "global progressive" control strategy with MYCIN-like confidences (Shortliffe, 1976), where "global" means that contextual information was used when needed.

A manual segmentation of the acoustic signal into phones and a description of the acoustic events for each segment were provided as input to the system. The system was evaluated on a test set of 100 French sentences, spoken by one male speaker. Each sentence included expressions made up from 13 phonemes, representing the "classical" manner of articulation phonetic classes. The test sentences were described to the system by graduate students who were unfamiliar with the details of the project. The system had an 83% top choice accuracy (including ties) and 94% accuracy for any position. The authors were encouraged by their results, both in terms of having developed a tool to

collect and evaluate acoustic-phonetic knowledge, and in the phoneme class recognition accuracy obtained.

While the previous attempts at building expert systems based on spectrogram reading have met with some success, I have reservations about the way in which the knowledge has been represented. Examples of the rules used in Johnson et al. (1984) and Stern et al. (1986) are quite specific with regard to the acoustic characteristics, such as numerical values for the formant frequency locations or frequency distributions of energy. I believe that the transformation from numerical measurements to qualitative descriptions should be separated from the inferences relating the acoustic characteristics to the phonemes. In addition, phonetic features, rather than phonemes should be deduced. This would enable the knowledge to be expressed more succinctly and to exploit the acoustic characteristics that phonemes with a given feature have in common.

5.1.3 Selection of a knowledge-based system shell

This research has focused on the acquisition and formalization of the knowledge base, rather than the development of a knowledge-based system, or shell, itself. As a result, existing technology has been used to implement a system for stop identification.

An initial implementation (Zue and Lamel, 1986) of a knowledge-base and a set of rules for stop identification used an available MYCIN-based (Shortliffe, 1975), backward-chaining system. Acoustic measurements were provided semi-automatically to the system and converted to qualitative descriptions. Rules related the qualitative descriptions to phonetic features, which were then mapped to phonemes. Beliefs (also called confidences, weights, or certainty factors) in the preconditions reflected uncertainty in the acoustic descriptions. Strengths in the rule conclusions reflected how strongly a given acoustic description indicated a phonetic feature. The MYCIN-based system had a very simple goal-directed control strategy. It set off to determine the identity of the stop, and in the process pursued the subgoals of deducing the voicing and place characteristics of the stop. In each case, the system exhaustively fired all pertinent rules. The control strategy could be modified somewhat by including preconditions to inhibit certain rules from firing.

The system (SS-1) was evaluated on 400 word-initial, intervocalic stops extracted from continuous speech. Table 5.1 compares the system performance to the performance of

Table 5.1: Comparison of human and SS-1 system identification performance.

condition		Number of tokens	Top choice(%)	Top 2 choice(%)
set 1	human(2)	200	90	92
	system	100	88	95
set 2	human(3)	200	92	96
	system	100	84	92
set 3	system	200	83	94

human spectrogram readers on two sets of 100 stops. The averaged human performance of 2 and 3 readers is given, for sets 1 and 2, respectively. The tokens in set 1 were also used to tune the system. System tuning involved setting the thresholds for the mapping functions, and refining the selected acoustic descriptions and the rules. For set 1, the system's performance was comparable to that of the experts. The performance of the system degraded by 4% when it was confronted with new data (set 2), whereas the experts' performance remained high. The degradation of performance from tuning to test data was attributed primarily to the "lack of experience" of the system; it had not learned all the acoustic descriptions and rules used by the experts. The system had comparable performance on another test set of 200 samples (set 3).

If performance in terms of recognition accuracy was the main objective, the SS-1 system may have been acceptable. However, an important objective of this research was to develop a system that models the problem-solving procedures used by human experts, something that the SS-1 system did not do very well. This was partly due to limitations imposed by the structure of the MYCIN-based system. The goal-directed inferencing of MYCIN did not enable the system to evaluate multiple hypothesis at any given time. In contrast, experts tend to use forward induction, and to simultaneously consider a set of possible candidates, although they may use goal-directed reasoning to confirm or rule out candidates. Since there is redundancy in the acoustic characteristics for a given phonetic feature, often only a subset of acoustic characteristics are needed to specify it. The goal-directed control structure of MYCIN always exhaustively fires all rules, while experts may quit when they have enough evidence for a feature. Other problems occurred with representing our knowledge in MYCIN's data structure, the "context-tree." The MYCIN system did not allow nodes at the same level of the context-tree to share information,

which made it difficult to model coarticulatory effects. As a result, it would have been difficult to increase the capabilities of the system to identify stops in other environments, such as consonant clusters.

Our experience with the SS-1 system indicated the need for a control strategy which better models the reasoning of spectrogram readers. The expert system shell ART, a commercial product developed by Inference Co., was selected because it integrates forward and backward reasoning, allows hypothetical reasoning and has "schemata" data-structures which provide frame-like capabilities. In addition, ART can be augmented to handle confidences in the preconditions and conclusions.

5.2 Knowledge acquisition

The knowledge incorporated in the implementation was obtained primarily by observing others reading spectrograms, by reading spectrograms myself, and by introspection. Using knowledge about the articulation of speech (and of stop consonants in particular) as a foundation, spectrograms of stop consonants were studied in an attempt to define acoustic correlates of their place of articulation and voicing characteristic. I also tried to determine how the acoustic evidence was weighed and combined in reaching a decision. The knowledge was tested by identifying unknown stop consonants in a limited context, as used in the listening and spectrogram reading experiments of Chapters 3 and 4. Trying to understand the errors that I made, led to changes and refinements in the decision strategy and to the inclusion of additional acoustic attributes.

Over the extent of this thesis work, I was also fortunate to be involved in attending and leading several spectrogram reading groups. Spectrogram reading sessions provide a unique opportunity to gather knowledge. All readers participate in the interpretation of the spectrogram, generally taking turns at identifying one or a few segments. When leading sessions of beginning spectrogram readers, we usually try to have them identify easy sounds first (such as strong fricatives, /r/'s, and other sounds with easily recognized acoustic correlates), leaving the more difficult interpretations until the end, when more contextual constraints can be applied. As the spectrogram readers gain experience, the spectrogram tends to be read from left-to-right, occasionally skipping over and returning to difficult regions. At his/her turn, each reader proposes a label or sequence of labels

for the region, and provides an explanation for his/her decision. When other readers disagree, there can be extensive discussion as to possible consistent interpretations. At the sessions, particular attention was paid to the acoustic attributes used by the readers and to the reasons they gave to justify their interpretation. Some of the sessions were tape recorded for further analysis.

Additional knowledge came from the spectrogram reading experiments discussed in Chapter 4 and from system development. By analyzing the errors made by the expert spectrogram readers, I was able to assess some of the tradeoffs they made. For example, some readers tended to favor the information provided by the burst location over that of the formant transitions. Other readers varied their strategy depending upon which information they felt was more robust in the given case. Each error was discussed with the reader who made it in order to elucidate the reader's reasoning. Implementing the system led to changes and refinements in the rules, particularly in the rule ordering. Rule development is an iterative, interactive process. Typically, a few examples were run through the system and, as a result, rules and rule interactions were modified.

5.3 Representation

This section describes the representation that has been developed for use in phonetic decoding. The representation combines knowledge from the acoustic theory of speech production (cf. Fant, 1960; Flanagan, 1972) and distinctive feature theory (Jacobson, Fant, and Halle, 1952; Chomsky and Halle, 1968).

5.3.1 Static knowledge base

Conceptually there are four levels of representation; phonemes (P), phonetic features (F), qualitative acoustic attributes (QAA) and acoustic measures (M). A block diagram of the representation is given in Figure 5.1. Moving from left-to-right in the figure provides a top-down description of the knowledge. Phonemes are defined in terms of their phonetic features. Internally, phonemes are also grouped into classes reflecting their manner of articulation, such as stops, vowels and fricatives. Grouping phonemes into classes allows some of the rules of the system to be expressed more succinctly. For example, the features

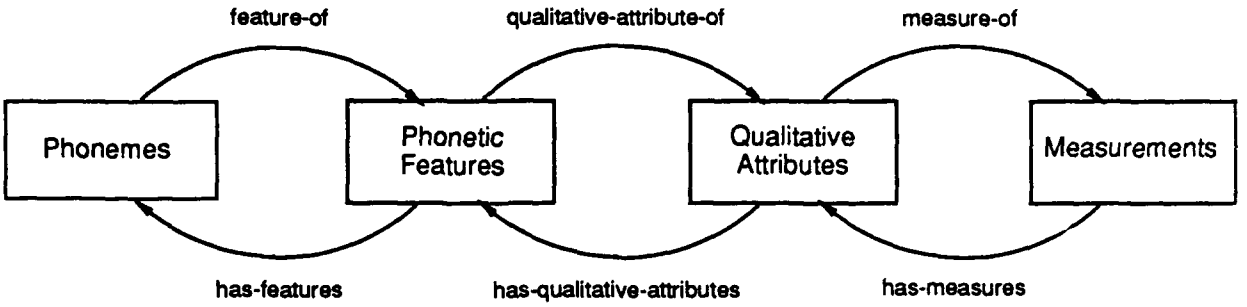


Figure 5.1: Knowledge representation.

[+ obstruent, - continuant] are associated with the class of stops, and inherited by each member of that class. The phonetic features are related to a set of acoustic attributes, each of which takes on a qualitative value. The qualitative attributes describe acoustic events in the speech signal and the canonical temporal and spectral characteristics of the features. Many of the qualitative attributes are based on our knowledge of the articulation of speech. These are either events seen in a spectrogram or derived from a quantitative acoustic measurement made in the speech signal. The links between the boxes mnemonically reflect their relationships.

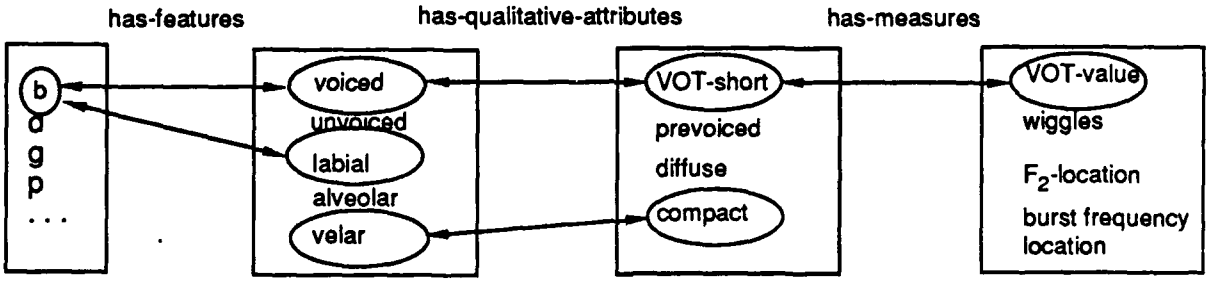


Figure 5.2: Subset of the knowledge used to represent stops. Some illustrative connections are indicated.

Figure 5.2 shows a subset of the knowledge used to represent the class of stop consonants. A stop is one of /b,d,g,p,t,k/. The stops are represented by their place of articulation and

their voicing characteristic.[1] The voicing characteristic of a stop is determined primarily by the acoustic attributes of voice-onset-time (VOT), prevoicing, and aspiration. The place of articulation decision is based on acoustic attributes describing the frequency and time distribution of the burst, the aspiration (if the stop is aspirated) and the formant transitions into the surrounding vowels. The acoustic attributes take on qualitative values, each of which is associated with an acoustic measure. For example, the VOT is the time measured from the release of the stop to the onset of voicing in the vowel. A VOT of 25 ms would be mapped into VOT-short. Similarly, the distribution of energy across frequency at the stop release may be characterized as compact, diffuse, even, or bimodal. An energy distribution that is compact has the energy of the release primarily located in a frequency range of 1-2 kHz.

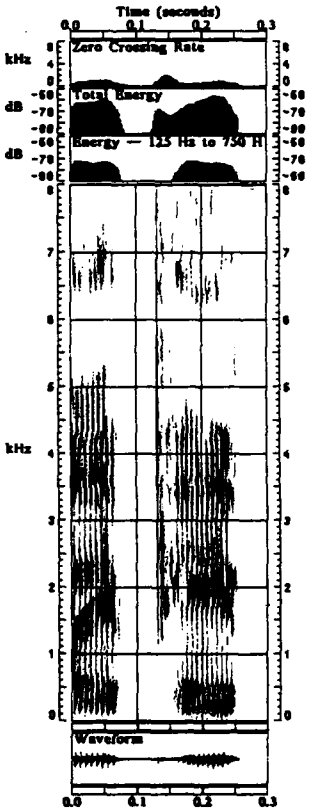
Vowels are also represented in the structure. The place of articulation of the vowel is determined by the tongue height and tongue position, and the position of the lips. The qualitative acoustic attributes associated with vowels describe the locations and movements of the formants. Acoustically, vowels are also described in terms of duration, which may be related to the tense/lax feature. The acoustic measures are the formant frequencies and the duration. For example, a back vowel has a high F_1 location and a low F_2 location, and an F_1 of 800 Hz is mapped to a high F_1. Semivowels, nasals and fricatives are represented analogously. Some of the place of articulation attributes for the fricatives and nasals are shared with the stops.

5.3.2 Dynamic knowledge base

In the preceding section the relationships between objects in the knowledge base were outlined. The knowledge is static in that it defines prototypes that do not change as a function of the rules. The representation of each stop can be thought of as a frame (Minsky, 1975), where the knowledge in the static database defines prototypes and the default values for each slot. A dynamic database of facts is created for each token as it is identified. The token exhibits specific acoustic characteristics which are converted to qualitative acoustic attributes. In turn, these qualitative acoustic attributes are used in phonetic decoding. The acoustic measures and qualitative acoustic attributes are

[1]The phonetic features used to describe the stops may be mapped into the distinctive features of Jacobson et al. (1952).

obtained from the utterance or by querying the "user." The responses satisfy the preconditions of rules enabling them to fire, resulting in deductions and further queries for additional information. The framework allows the queries to be replaced with function calls to measure parameters directly or with database requests to retrieve prestored measures and/or attributes.



Left vowel: front, high
duration not-long, not-short
F_2 falling, amount large
F_3 falling

Stop: closure not-prevoiced
release VOT not-short, not-long
aspirated
aspiration falling into release
burst-location broad
energy-distribution even
strength weak

Right vowel: front, high
duration not-short, not-long
F1 not-falling
F2 not-falling, not-rising
F3 not-rising, maybe-falling (during aspiration)

Figure 5.3: Facts in the dynamic database for the token /ɪpɪ/ .

Figure 5.3 shows a spectrogram of /ɪpɪ/ and the acoustic attributes associated with the token. The acoustic attributes were determined by querying the user or from the phonetic transcription. Rules scan the database of facts to determine the stop's identity.

5.3.3 Probing the knowledge base

Some facilities have been developed for probing both the static and dynamic knowledge bases. A *what-is* or *what-has* question returns a table look-up or definitional answer. A

Table 5.2: Examples of the types of queries recognized by the system.

Question	Object	Answer
what-is	phoneme	feature bundle
	feature	set of acoustic attributes
	acoustic attribute	description (value in context)
what-has	feature(s)	phonemes having feature(s)
	acoustic attribute	features having QAA
why	phoneme	associated deduced features
	feature	associated QAA's
why-not	phoneme	missing features
	feature	missing or contradictory QAA

why or *why-not* question is generally used to justify in a specific example. Some examples of the types of queries and the forms of responses are given in Table 5.2.

The system response to the query *"what-is a /p/?"* is that a /p/ is a voiceless, labial stop. The response to *"what-is voiced?"* is a list of all the QAA's that specify voiced: short-VOT, prevoiced, and not-aspirated. The answer to the query *"what-has the feature voiced?"* is the set of stops /b,d,g/. The *why-not* query is used to ask the system why a deduction was not made. The system responds with a list of missing and/or contradictory information.

5.4 Qualitative acoustic attributes

The qualitative acoustic attributes (QAA's) describe the acoustic events visible in the spectrogram. Each segment is represented by a set of QAA's specific for the type of segment. Table 5.3 lists some examples of qualitative acoustic attributes used to describe the stop consonants. Each QAA is used to determine the place or the voicing of the stop. A complete listing of the qualitative attributes used in the implementation, along with an example of each, is given in Appendix C. The stop in Figure 5.3 has the qualitative acoustic attributes listed in the figure.

QAA's are obtained by querying the user or by mapping acoustic measures. Certain combinations of qualitative acoustic attributes cannot co-occur. For example, it would

Table 5.3: Examples of qualitative acoustic attributes of stops.

dimension	region	attribute
voicing	release	VOT-short
		VOT-long
		aspirated
	closure	prevoiced
place	release	burst-location-HF
		burst-location-MF
		burst-location-LF
		burst-location-bimodal
		energy-distribution-diffuse
		energy-distribution-compact
		energy-distribution-even
		energy-distribution-bimodal
		strength-strong
		strength-weak

be meaningless to have a burst-strength that was both weak and strong. To prevent such a situation, the rules that query the user for information take account of the facts already known. For example, if the user responds that burst-strength is strong, then the system will not query to determine if the burst-strength is weak, but instead automatically asserts that the burst-strength is not-weak.

5.5 Rules and strategy

Plausible strategies for some of the cognitive aspects of spectrogram reading are simulated through the rules. While I am unable to verify that spectrogram readers use these or similar strategies, the strategies "feel right" to the expert. Much of the reasoning is data-driven—the reader sees acoustic events in the spectrogram and makes deductions based on them. The reader is able to combine multiple cues in forming a judgement and to consider multiple hypotheses at once. The reader may use goal-directed reasoning to confirm or rule out hypotheses. Readers are also able to deal with uncertainty in the acoustic evidence and, to some degree, with acoustic information that may be contradictory. In order to rank competing hypotheses, readers somehow weigh the evidence and form a decision.

An attempt has been made to capture most of the above cognitive aspects in the implementation. The implementation integrates data-driven and goal-directed reasoning. The data-driven rules make deductions based on the qualitative acoustic attributes. Goal-directed reasoning is used to query the user (or database) for new information and to confirm or rule out hypotheses. The system models the human capability to simultaneously consider multiple hypotheses by maintaining a ranking of all candidates at all times. The rules may be one-to-one, as linking phonetic features and phonemes, or one-to-many and many-to-one, as in deducing phonetic features from qualitative acoustic attributes. Thus, the rules provide the capability to handle the problems of multiple cues and multiple causes. How readers actually combine information from multiple sources and deal with uncertain evidence has not been determined; however, what the reader appears (or claims) to be doing is modeled. Several strategies for combining evidence have been investigated and are discussed in section 5.6. Uncertainty in the acoustic evidence is modeled by allowing users to specify that an acoustic attribute is present, absent, or maybe present. Constraining the system to use uncertain acoustic attributes only after using definitive ones provides a mechanism for relaxing constraints under uncertainty. Uncertainty in the deductions is handled by associating a strength with each deduction.

5.5.1 Rules

In this section the different types of rules used in the system are described, and examples are provided. Appendix D contains a listing of the rules. As mentioned in section 5.3, different rule sets cover the relations between the levels in the representation. Rules map phonetic features to phonemes, relate qualitative acoustic attributes to phonetic features, and map acoustic measurements to qualitative acoustic attributes.

Table 5.4: Phonetic features of stops.

	b	d	g	p	t	k
voiced	+	+	+	-	-	-
labial	+	-	-	+	-	-
alveolar	-	+	-	-	+	-
velar	-	-	+	-	-	+

Definitional rules: A set of "definitional rules" map the phonemes to their phonetic features. The representation of stops according to their place of articulation and their voicing characteristic is shown in Table 5.4. The rules encode the information in the table explicitly. An example of a definitional rule is:

> **If** the voicing of the stop is *voiced,*
> **and** the place of articulation of the stop is *alveolar,*
> **then** the identity of the stop is */d/*.

While conceptually there are different definitional rules for each stop, they are all actually implemented with one rule. The rule also combines the beliefs associated with each feature and provides a belief in the identity of the stop. The use of beliefs is discussed further in section 5.6. The following rule explicitly captures the knowledge that a stop can be described in terms of its voicing characteristic and its place of articulation.

> **If** the voicing of the unknown-stop is *voicing-value* with *voicing-belief*
> **and** the place of articulation of the unknown-stop is *place-value* with *place-belief*
> **and** there exists a prototype stop with identity *identity*
> **and** with voicing *voicing-value*
> **and** with place of articulation *place-value*
> **then** the identity of the unknown-stop is *identity* with belief(*voicing-belief,place-belief*).

Rules relating qualitative acoustic attributes to features: The relationships between the qualitative acoustic attributes and the phonetic features are complicated. The majority of the rules in the implementation deal with these relationships. The rules are all of the form:

> **If** precondition(s)
> **then** conclusion(s) with strength(s)

The preconditions are generally facts that exist in the database. However, the absence of a fact may also be used as a precondition: whenever the fact exists, it serves to inhibit the rule from firing. In order for a rule to "fire," all of its preconditions must be met. A rule is said to have "fired" when all of its actions have been taken. The actions typically modify the dynamic database.

A given phonetic feature may be signalled by several qualitative acoustic attributes, resulting in multiple rules to deduce the phonetic feature. For example, both a long-VOT and the presence of aspiration in the stop release are cues for the feature voiceless. The corresponding two rules are:

If the VOT is *long*,
then there is *strong evidence* that voicing characteristic is *voiceless* .

If the release is *aspirated*,
then there is *strong evidence* that voicing characteristic is *voiceless* .

If, as in the example of Figure 5.4(c), the preconditions of both of the rules are satisfied, then the belief that the voicing of the stop is voiceless will be quite strong. (The way in which the evidences are combined is discussed in section 5.6.) Not all the qualitative acoustic attributes for a phonetic feature are always present. For any particular acoustic segment, some or all of the rules may have their preconditions satisfied and those rules will fire.

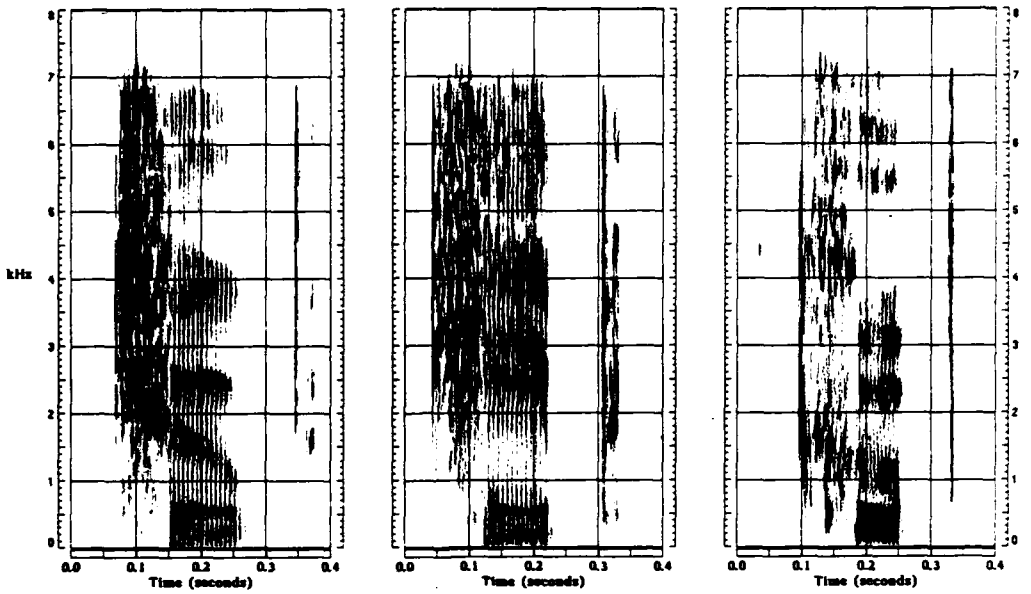
A given qualitative acoustic attribute may be indicative of different phonetic events, resulting in rules that have multiple deductions with respect to the context. For example, in the absence of contextual information, a burst spectrum that has energy predominantly at high frequencies is likely to indicate an alveolar place of articulation. However, if the stop is in a syllable with a front vowel, the stop is also likely to be velar and may be labial. The contextual influences are directly incorporated into the rules as follows:

If the burst-location is *high-frequency*,
then there is *strong evidence* that the place of articulation is *alveolar* .

If the burst-location is *high-frequency*,
and the vowel is *front*
then there is *strong evidence* that the place of articulation is *velar*
and there is *weak evidence* that the place of articulation is *labial* .

Figure 5.4 illustrates an example requiring the use of such contextual information. The spectral characteristics of the stop release in the left and middle spectrograms are quite similar: they both have a predominance of high frequency energy. In this example, it would not be easy to determine the identity of either stop only by visual inspection of the release. The spectral characteristics of the release are consistent with both a /t/ and a front-/k/. However, knowledge that the following vowel in the left spectrogram is an /u/ indicates that the stop is a /t/. The spectral characteristics of a back, rounded /k/ in the syllable /ku/ are quite different, as can be seen in the right spectrogram in Figure 5.4.

The presence or absence of acoustic evidence may be important. For example, if a stop



(a) /tup/ (b) /kip/ (c) /kup/

Figure 5.4: Spectrograms illustrating contextual variation.

is syllable-initial and has a VOT value that is medium (maybe-short and maybe-long), then the presence of aspiration indicates that it is voiceless, and the absence of aspiration is indicative of voiced. Two rules that use aspiration to deduce the voicing characteristic are:

> **If** the stop is syllable-initial
> **and** the VOT is *maybe-long* and *maybe-short*,
> **and** the release is *aspirated*,
> **then** there is *strong evidence* that voicing characteristic is *voiceless*.

> **If** the stop is syllable-initial
> **and** the VOT is *maybe-long* and *maybe-short*,
> **and** the release is *not-aspirated*,
> **then** there is *medium evidence* that voicing characteristic is *voiced*.

Note that the presence of aspiration is a stronger indicator of voiceless than the lack of aspiration is of voiced. In other cases, the presence of an acoustic attribute may indicate a feature, but the absence of the acoustic attribute does not provide negative evidence for that feature. One such acoustic attribute is a double burst (see Figure 1.2). When a double burst is observed it is a strong indicator of a velar place of articulation. However,

since a double burst is not that common, the system must have some evidence that the place of articulation is velar before attempting to find a double burst. The double-burst rule is:

> **If** the place of articulation is *velar* with *belief*
> **and** the release has a *double burst*
> **then** there is *strong evidence* that place of articulation is *velar*.

The value of the voicing characteristic and of the place of articulation are deduced independently. While it is possible to write rules to deduce phonemes directly instead of features, I have chosen not to do so. Deducing phonetic features rather than phonemes adds another level of representation and generalization. This allows commonality in the rules for place or voicing to extend to different manner classes. For example, vowels and nasals are both shorter preceding voiceless consonants than voiced consonants in the same syllable (House and Fairbanks, 1953; Peterson and Lehiste, 1960; Raphael et al., 1975; Klatt, 1976; Hogan and Rozsypzl, 1980; Glass, 1983; Zue and Sia, 1984). This phonological effect can be captured in one rule, instead of individually for each phoneme. The formant motion between a vowel and a consonant depends primarily on the place of articulation of the consonant, and not on its identity. Thus, for example, the qualitative acoustic attribute of falling formants can be associated with the feature labial, covering multiple phonemes.

Phonotactic constraints are implemented in rules which account for the phonetic context. For example, if the stop is preceded by a fricative, the system will attempt to determine whether the fricative is an /s/ or a /z/. If the fricative is a /z/, the system asserts that the stop is syllable-initial. If the fricative is an /s/, the system must determine whether of not the /s/ and the stop form a cluster. If the stop is in a cluster with the /s/, then the stop is voiceless. If the stop is not in a cluster with the /s/, then there is a syllable boundary between the fricative and the stop, and the syllable-initial rules to determine the voicing of the stop may be applied.

The context is specified in the preconditions of the rules to ensure that they fire only under the appropriate conditions. When the stop is preceded by a fricative, the formant motion in the vowel to the left of the fricative is not used, since the formant transitions should always indicate the alveolar place of articulation of the fricative. This is implemented by preconditions in the vowel formant rules which specify that the right context cannot be a fricative. For a stop preceded by a homorganic nasal, the formant motion in the vowel

preceding the nasal is used to infer the place of articulation of the nasal, which is the same as that of the stop.

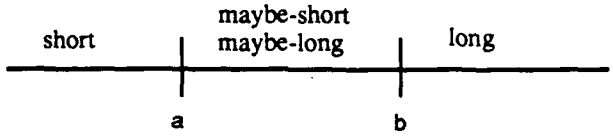


Figure 5.5: Example of mapping ranges for numerical quantities.

Mapping rules: The mapping rules convert acoustic measurements into qualitative attributes. The mapping rules are all implemented as backward-chaining rules, and therefore do not fire unless they are called upon to produce a result needed by another rule. The mappings are schematically illustrated in Figure 5.5. The rules which map from numerical quantities into qualitative acoustic attributes are of the form:

If the measured-value is $< a$
then the attribute has the qualitative-value *short*
else if the measured-value is $> b$
    then the attribute has the qualitative-value *long*
otherwise the attribute has the qualitative-values *maybe-short* and *maybe-long*

The mapping rules typically divide the range into disjoint regions, where measures falling between regions are associated with both labels. The mapping ranges were hand-selected by looking at histograms of the measure on a set of training samples. However, these could be statistically trained if enough data were analyzed.

5.5.2 Control strategy

Spectrogram readers appear to extract acoustic attributes in the spectrogram and to propose a set of features consistent with the attributes. The candidate set is refined by looking for additional acoustic evidence to confirm or rule out some of the possibilities. The control strategy attempts to model the behavior of spectrogram readers. The order in which the rules fire can be controlled by priorities associated with the rules and by the use of preconditions. Both of these methods are used to affect the behavior of the system so as to have it appear more "intelligent."

The system uses general information before using more specific information. This is implemented by associating higher priorities with the more general rules. Preconditions are also used to prevent the system from asking for detailed information too early. An example was shown in the double-burst rule, where there had to be some belief that the place of articulation was velar before the rule could be applied.

The system maintains a ranking of all candidates. Each time a new fact is asserted into the database, the ranking rules reorder the candidates for place and voicing. In this way, the system simultaneously considers multiple hypotheses at the same time. The list of ordered candidates enables the system to pursue the most likely candidate first. The rules are ordered so as to seem logical to the user. For example, if labial is the top candidate, the system tries to determine if the stop has a weak release, as a weak release provides confirming evidence for labial. If the top candidate is alveolar, and the second candidate is labial, the system will attempt to find out if the release is strong, as a strong release favors alveolar over labial. However, if the top two candidates are alveolar and velar, rules using the strength of the release are postponed, since the release strength is not a good attribute to distinguish between them.

Without specific "termination" rules, the system exhaustively fires all rules until there are no more left. However, the system may be run in a mode where, when it has enough evidence for a feature, it does not exhaustively pursue all the alternatives. This behavior is implemented by rules which attempt to confirm the top candidate and to rule out the closest competitor when the belief in the top candidate is large enough. If the belief in the top candidate (and the distance between the top two candidates) increases, then the system confirms the top candidate and no longer attempts to determine the value of the feature.

The behavior of the system is apparent to the user primarily through the rules which query the user.[2] All of the queries are the result of backward chaining and are implemented in ART using "goals." In this way, the system does not ask for information that it does not need. When a rule needs some information in order to fire, and there exists a backward chaining rule that can supply the information, a "request" is made. The rule with the highest priority that can supply the requested information will be fired, resulting in a query to the user. The order in which the system queries the user for information

[2]The term "user" refers to all queries. In reality, the query might be made to a database of facts or to make measures in the acoustic signal directly.

depends on what facts are in the database. Using preconditions and priorities to affect the control, the system can ask questions that pursue the best candidate.

The queries made of the user fall into three categories: numbers corresponding to an acoustic measurement such as a duration or formant frequency; classifications such as the segment's identity, manner class or syllable-position; and qualitative acoustic attributes. The QAA's are categorical: the user is queried as to the presence of the QAA, and may respond with *yes*, *no*, *maybe*, or *can't-tell*. The response of *maybe* handles (to a limited degree) uncertainty in the acoustic evidence. *Can't-tell* is only used when the acoustic evidence needed to answer the query is not visible in the spectrogram. For example, if the system is asking whether F_3 is falling, but F_3 is not visible in the spectrogram, then that information can not be supplied.

5.5.3 An example of identifying a stop

In this section, the stop in Figure 5.3 is identified by the rule-based system.

```
; Each rule firing is preceded by FIRE n, where n is the nth rule to fire
; fact numbers (f- ) and goal numbers (g- ) used to trigger rule are shown in parentheses
; Query rules provide the questions asked of the user with allowable responses:
;       (yes maybe no cant-tell)
; ==> are asserted facts
; <== are retracted facts

; set up token to be identified, it is a both a token and a stop
FIRE 1 TOKEN-TO-BE-IDENTIFIED ()
==> f-1 [INSTANCE-OF T1 TOKEN]
==> f-2 [INSTANCE-OF T1 STOP]

; a stop has intervals for the closure, release and aspiration
FIRE 2 CREATE-INTERVALS-FOR-STOP (f-2)
==> f-3 [INSTANCE-OF CL01 CLOSURE]
==> f-4 [INTERVAL-OF CL01 T1]
==> f-5 [INSTANCE-OF REL1 RELEASE]
==> f-6 [INTERVAL-OF REL1 T1]
==> f-7 [INSTANCE-OF ASP1 ASPIRATION]
==> f-8 [INTERVAL-OF ASP1 T1]

; R1 is to the right of T1
FIRE 3 SETUP-RIGHT-CONTEXT (f-1)
==> f-9 [INSTANCE-OF R1 TOKEN]
==> f-10 [RIGHT-OF R1 T1]
```

==> f-11 [CLASS R1 VOWEL]
==> f-12 [VOICING-CHARACTERISTIC R1 F-VOICED]

; L1 is to the left of T1
FIRE 4 SETUP-LEFT-CONTEXT (f-1)
==> f-13 [INSTANCE-OF L1 TOKEN]
==> f-14 [LEFT-OF L1 T1]
==> f-15 [CLASS L1 VOWEL]
==> f-16 [VOICING-CHARACTERISTIC L1 F-VOICED]

; look up the syllable position of the stop
FIRE 5 DETERMINE-SYLLABLE-POSITION-VOWEL (f-2 f-9 f-10 f-11 f-13 f-14 f-15)
==> f-17 [PROPERTY-OF SYLLABLE-INITIAL T1]

FIRE 6 QUERY-BURST-VISIBLE (f-2 f-5 f-6 g-1)
Can you see a stop release in the spectrogram?
==> f-18 [QUALITATIVE-ATTRIBUTE-OF Q-BURST-VISIBLE-YES REL1]

FIRE 7 QUERY-MVOT-INITIAL (f-2 f-5 f-6 f-17 g-2)
What is the VOT value [in ms]?
==> f-19 [M-VOT REL1 30.6]

; the VOT of 30.6 ms value is neither short nor long
FIRE 8 MAP-VOT-disjunct-3 (f-2 f-5 f-6 f-17 f-18 f-19 g-3)
==> f-20 [QUALITATIVE-ATTRIBUTE-OF Q-VOT-SHORT-NO REL1]
==> f-21 [QUALITATIVE-ATTRIBUTE-OF Q-VOT-LONG-NO REL1]

FIRE 9 QUERY-PREVOICING (f-2 f-3 f-4 f-13 f-14 f-15 g-4)
Is there prevoicing during the closure interval?
==> f-22 [QUALITATIVE-ATTRIBUTE-OF Q-PREVOICED-NO CL01]

; VOT short-no, long-no and prevoiced-no → voiceless
FIRE 10 VOICING-PREVOICING-VOICELESS-MEDIUM (f-2 f-3 f-4 f-5 f-6 f-15 f-16 f-20 f-21 f-22)
==> f-23 [ADD-TO-SCORE 1 T1 VOICING-CHARACTERISTIC F-VOICELESS MEDIUM-EVIDENCE]

; the next three rules update the scores to maintain a ranking of all candidates,
; later updates are skipped
FIRE 11 COPY-FACT (f-23)
==> f-24 [COPY-ADD-TO-SCORE T1 VOICING-CHARACTERISTIC F-VOICELESS MEDIUM-EVIDENCE]

FIRE 12 ADD-TO-SCORE (f-24)
<== f-24 [COPY-ADD-TO-SCORE T1 VOICING-CHARACTERISTIC F-VOICELESS MEDIUM-EVIDENCE]

==> f-25 [SCORE VOICING-CHARACTERISTIC T1 F-VOICELESSS 0.5]
==> f-26 [ADDED-TO-SCORE T1 VOICING-CHARACTERISTIC F-VOICELESS MEDIUM-EVIDENCE]

FIRE 13 RANK-VOICING (f-2 f-25)
voicing order for T1: f-voiceless 0.5

f-voiced 0
==> f-27 [TOP-CANDIDATE VOICING-CHARACTERISTIC T1 F-VOICELESS 0.5]

FIRE 14 QUERY-ASPIRATION-VOICELESS (f-2 f-5 f-6 f-7 f-8 g-5)
Is the release aspirated?
==> f-28 [QUALITATIVE-ATTRIBUTE-OF Q-ASPIRATED-YES ASP1]

; VOT short-no, long-no and aspirated-yes → voiceless
FIRE 15 VOICING-ASPIRATION-YES-VOICELESS (f-2 f-5 f-6 f-7 f-8 f-17 f-20 f-21 f-27 f-28)
==> f-29 [ADD-TO-SCORE T1 VOICING-CHARACTERISTIC F-VOICELESS STRONG-EVIDENCE]

FIRE 20 RANK-VOICING-CONFIRM (f-1 f-29)
voicing order for T1: f-voiceless 1.3
f-voiced 0
==> f-30 [CONFIRM VOICING-CHARACTERISTIC T1 F-VOICELESS 1.3]

FIRE 21 QUERY-BURST-LOCATION-LF (f-2 f-5 f-6 g-6)
Is the burst (including frication noise) primarily at low frequency [roughly below 2 kHz]?
==> f-31 [QUALITATIVE-ATTRIBUTE-OF Q-BURST-LOCATION-LF-NO REL1]

FIRE 22 QUERY-BURST-LOCATION-MF (f-2 f-5 f-6 g-7)
Is the burst (including frication noise) primarily at mid frequency [roughly 2-4 kHz]?
==> f-32 [QUALITATIVE-ATTRIBUTE-OF Q-BURST-LOCATION-MF-NO REL1]

FIRE 23 QUERY-BURST-LOCATION-HF (f-2 f-5 f-6 g-8)
Is the burst (including frication noise) primarily at high frequency [mainly above 4 kHz]?
==> f-33 [QUALITATIVE-ATTRIBUTE-OF Q-BURST-LOCATION-HF-NO REL1]

FIRE 24 QUERY-BURST-LOCATION-BROAD (f-2 f-5 f-6 g-9)
Is the burst (including frication noise) evenly distributed across all frequencies [0-8kHz]?
==> f-34 [QUALITATIVE-ATTRIBUTE-OF Q-BURST-LOCATION-BROAD-YES REL1]

; broad burst location → labial
FIRE 25 PLACE-BURST-BROAD-INITIAL (f-2 f-5 f-6 f-17 f-34)
==> f-35 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL STRONG-EVIDENCE]

FIRE 30 RANK-PLACE (f-2 f-35)
place order for T1: F-LABIAL 0.8
f-alveolar 0
f-velar 0
==> f-36 [TOP-CANDIDATE PLACE-OF-ARTICULATION T1 F-LABIAL 0.8]
==> f-37 [2ND-CANDIDATE PLACE-OF-ARTICULATION T1 F-ALVEOLAR 0]

FIRE 31 QUERY-TOP-BURST-STRENGTH-WEAK (f-2 f-5 f-6 f-18 f-36 g-10)
Is the burst weak relative to the vowel?
==> f-38 [QUALITATIVE-ATTRIBUTE-OF Q-STRENGTH-WEAK-YES REL1]

; release weak → labial
FIRE 32 PLACE-STRENGTH-WEAK-INITIAL (f-2 f-5 f-6 f-17 f-38)
==> f-39 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL MEDIUM-EVIDENCE]

```
FIRE 37 RANK-PLACE (f-2 f-36 f-37)
   place order for T1: f-labial 1.3
                       f-alveolar 0
                       f-velar 0
==> f-40 [TOP-CANDIDATE PLACE-OF-ARTICULATION T1 F-LABIAL 1.3]
==> f-41 [2ND-CANDIDATE PLACE-OF-ARTICULATION T1 F-ALVEOLAR 0]

FIRE 38 QUERY-F2-FALLING (f-2 f-9 f-10 f-11 f-40 g-11)
   Is F2 falling into the stop? [Pay attention to the 30-40 ms near the stop]
==> f-42 [QUALITATIVE-ATTRIBUTE-OF Q-F2-FALLING-NO R1]

FIRE 39 QUERY-F2-FALLING (f-2 f-13 f-14 f-15 f-40 g-12)
   Is F2 falling into the stop? [Pay attention to the 30-40 ms near the stop]
==> f-43 [QUALITATIVE-ATTRIBUTE-OF Q-F2-FALLING-YES L1]

FIRE 40 QUERY-F2-MOTION-AMT-LARGE (f-13 f-14 f-15 f-43 g-13)
   Does F2 move a large amount (> 200 Hz) from the stop to the vowel?
==> f-44 [QUALITATIVE-ATTRIBUTE-OF Q-F2-AMT-LARGE-YES L1]

; look up vowel identity
FIRE 41 QUERY-IDENTITY (f-13 f-14 f-15 g-14)
==> f-45 [ASKED-IDENTITY L1 /ɪ/]

; map from vowel identity to features
FIRE 42 VOWEL-FEATURES (f-45)
==> f-46 [HAS-FEATURES L1 F-HIGH]
==> f-47 [HAS-FEATURES L1 F-FRONT]
==> f-48 [HAS-FEATURES L1 F-TENSE]
==> f-49 [HAS-FEATURES L1 F-LONG]

; left F2 falling a large amount into stop, front vowel → labial, alveolar, not velar
FIRE 43 FORMANTS-LF2-FALLING-LARGE-disjunct-1 (f-2 f-13 f-14 f-15 f-43 f-44 f-47)
==> f-50 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL STRONG-EVIDENCE]
==> f-51 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-ALVEOLAR WEAK-EVIDENCE]
==> f-52 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-VELAR MEDIUM-NEGATIVE-EVIDENCE]


FIRE 48 RANK-PLACE-CONFIRM (f-2 f-40 f-41 f-50 f-51 f-52)
   place order for T1: f-labial 2.1
                       f-alveolar 0.2
                       f-velar -0.5
==> f-53 [CONFIRM PLACE-OF-ARTICULATION T1 F-LABIAL 2.1]
==> f-54 [RULEOUT PLACE-OF-ARTICULATION T1 F-ALVEOLAR 0.2]

FIRE 49 QUERY-F2-RISING (f-9 f-10 f-11 f-54 g-15)
   Is F2 rising into the stop? [Pay attention to the 30-17 ms near the stop]
==> f-55 [QUALITATIVE-ATTRIBUTE-OF Q-F2-RISING-NO R1]
```

; look up vowel identity
FIRE 50 QUERY-IDENTITY (f-9 f-10 f-11 g-15)
==> f-56 [ASKED-IDENTITY R1 /ɪ/]

; map from vowel identity to features
FIRE 51 VOWEL-FEATURES (f-56)
==> f-57 [HAS-FEATURES R1 F-HIGH]
==> f-58 [HAS-FEATURES R1 F-FRONT]
==> f-59 [HAS-FEATURES R1 F-TENSE]
==> f-60 [HAS-FEATURES R1 F-LONG]

; lack of formant motion → not labial
FIRE 52 FORMANTS-RF2-FLAT-disjunct-1 (f-2 f-9 f-10 f-11 f-54 f-55 f-58)
==> f-61 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL MEDIUM-NEGATIVE-EVIDENCE]

FIRE 57 RANK-PLACE (f-2 f-53 f-54 f-61)
place order for T1: f-labial 1.6
f-alveolar 0.2
f-velar -0.5
==> f-62 [TOP-CANDIDATE PLACE-OF-ARTICULATION T1 F-LABIAL 1.6]
==> f-63 [2ND-CANDIDATE PLACE-OF-ARTICULATION T1 F-ALVEOLAR 0.2]

FIRE 58 QUERY-CONFIRM-F3-FALLING (f-2 f-13 f-14 f-15 f-62 g-16)
Is F3 falling into the stop? [Pay attention to the 30-17 ms near the stop]
==> f-64 [QUALITATIVE-ATTRIBUTE-OF Q-F3-FALLING-YES L1]

; left F3 falling into stop, front vowel → labial
FIRE 59 FORMANTS-LF3-FALLING-disjunct-1 (f-2 f-13 f-14 f-15 f-47 f-64)
==> f-65 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL MEDIUM-EVIDENCE]

FIRE 64 RANK-PLACE-CONFIRM (f-2 f-62 f-63 f-65)
place order for T1: f-labial 2.1
f-alveolar 0.2
f-velar -0.5
==> f-66 [CONFIRM PLACE-OF-ARTICULATION T1 F-LABIAL 2.1]
==> f-67 [RULE-OUT PLACE-OF-ARTICULATION T1 F-ALVEOLAR 0.2]

FIRE 65 QUERY-CONFIRM-ASPIRATION-TAIL (f-2 f-7 f-8 f-28 f-66 g-18)
Does the low frequency edge of the aspiration lower in frequency into the stop?
==> f-68 [QUALITATIVE-ATTRIBUTE-OF Q-TAIL-YES ASP1]

; rising location of aspiration → labial
FIRE 66 ASPIRATION-LABIAL-TAIL (f-2 f-7 f-8 f-9 f-10 f-11 f-28 f-57 f-68)
==> f-69 [ADD-TO-SCORE T1 PLACE-OF-ARTICULATION F-LABIAL MEDIUM-EVIDENCE]

; confirmed place as labial
FIRE 71 RANK-PLACE-CONFIRM (f-2 f-66 f-67 f-69)
place order for T1: f-labial 2.6

```
                    f-alveolar 0.2
                    f-velar -0.5
==> f-70 [CONFIRMED PLACE-OF-ARTICULATION T1 F-LABIAL 2.6]

FIRE 72 IDENTIFY-STOP (f-1 f-2 f-30 f-70)
token: T1  stop: p    voicing: f-voiceless 1.3  place: f-labial 2.6
           stop: t    voicing: f-voiceless 1.3  place: f-alveolar 0.2
```

5.6 Scoring

The rules provide evidence that a given feature has a particular value. Since there are multiple rules which deduce the same feature, some way is needed to combine the evidence from the different rules. Combining evidence is an unsolved problem in expert systems research. There have been different approaches to the problem, including probabilistic, such as Bayesian (Duda et al., 1976), fuzzy logic (Zadeh, 1975), and more ad hoc formulations such as counting (Keeney and Raiffa, 1976) and the MYCIN-combine function (Shortliffe, 1975).

It is beyond the scope of this thesis to try to determine an optimum scoring strategy. However, there are some properties that a reasonable scoring scheme for this application should have:

- The scoring must be able to handle both positive and negative evidence.
- The combining of evidence should be order independent.
- The combining of evidence should be monotonic. Positive evidence can never decrease the belief in something and negative evidence can never increase it.
- Since the rules assert conclusions with strengths, the combining should also preserve the relative strengths of the conclusions. A weak conclusion cannot increase the belief more than a strong one can. The converse is also true. In addition, a strong positive conclusion and a weak negative conclusion cannot combine to reduce the belief in something.

The goal in building a knowledge-based system is to use domain knowledge to solve the problem. By using rules that are based on our knowledge about the articulation of speech

and our experience in spectrogram reading, the hope is that reasonable performance can be obtained with a small amount of training data. Much more training data would be required to train a statistical classifier on the qualitative acoustic attributes. A probabilistic scoring strategy (Duda et al., 1976; Duda et al., 1981) could also be used to model the correlations between the qualitative acoustic attributes and the phonetic features, if there was enough training data.

Two simple scoring schemes satisfying the above properties have been investigated. The first assigned numerical values to weak, medium, strong and certain evidence, and summed the evidence. Positive evidence was added, and negative evidence subtracted. The numbers used were:

with-certainty	=	1.0
strong-evidence	=	0.8
medium-evidence	=	0.5
weak-evidence	=	0.2

The second scheme counted the number of reasons of each strength. The candidates were ranked according to lexicographic ordering (Keeney and Raiffa, 1976). In lexicographic ordering the candidate with the largest number of strong reasons is the best. If two candidates have an equal number of strong reasons, they are ranked by the number of medium reasons, etc. In addition, no number of strong evidences can combine to be certain, no number of medium evidences can equal a strong, and no number of weak evidences can equal a medium.

5.7 Evaluation

The rule-based implementation has been evaluated in several ways. First, system performance on a subset of the data from each of the five tasks described in Chapter 2 (p. 20) is presented. Second, the rules used in the SS-1 system (Zue and Lamel, 1986) were reimplemented in ART for comparison, and the tokens in set 1 (Table 5.1) were used to assess the sensitivity of the system to the scoring method.

Evaluation on the five tasks: The system was tested on a subset of the tokens used to evaluate the spectrogram readers. The tokens for each task were divided into two sets. The first set contained tokens that were *heard correctly* by all listeners *and* were

read correctly by all readers (AC). The second set contained tokens that were *misheard or misread* by at least one subject (SE). The tokens were selected so as to include roughly equal numbers of each stop in a variety of vowel contexts. System performance for the five tasks is given in Table 5.5 for the two subsets. Both the top choice[3] and top-two choice accuracies are provided. The system performance was about the same for syllable-initial singleton stops and syllable-initial stops preceded by /s/ or /z/. The system identified singleton stops better when they occurred in syllable-initial position than in non-syllable-initial position. Performance was better for non-initial stops in nasal clusters than for singleton non-initial stops (compare tasks 4 and 5). The system failed to propose the correct candidate in 2% of the AC tokens and 12% of the SE tokens. As expected, the performance on the AC subset was better than on the tokens that were misidentified by humans.

For this evaluation, syllable position information was provided to the system. In task 2 the system was also told the identity of the fricative. When the fricative was an /s/, the system proposed two alternatives— one in which the /s/ and stop formed a cluster and the other in which they did not. If the system identified the stop correctly for the appropriate condition (cluster or not), the system was credited with identifying the stop correctly. In tasks 4 and 5, the system was informed that the stops were non-initial and, in task 5, that the preceding sonorant was a nasal forming a cluster with the stop.

Table 5.5: System evaluation on the five tasks.

	Percent correct: top/top 2			
	N	AC	N	SE
task 1	24	88/96	27	82/93
task 2	26	89/96	11	64/73
task 3	14	100/100	17	82/94
task 4	18	67/89	19	58/68
task 5	12	100/100	6	83/100

Analysis of errors on the AC tokens: Even though listeners and spectrogram readers were able to identify the tokens in the AC subset, the system still made errors

[3]When there was a tie for the top choice, the system was credited with having identified the stop correctly. Ties occurred on only 6 of the 174 tokens.

Table 5.6: Confusion matrices for system identification of AC and SE tokens.

AC answer	number tokens	percent correct	Listener's response b	d	ɡ	p	t	k	ǰ	č
b	14	93	13			1				
d	13	62	2	8	2		1			
ɡ	11	91	1		10					
p	16	94		1		15				
t	18	100					18			
k	18	78					4	14		
ǰ	2	100							2	
č	2	100								2

(a) AC tokens

SE answer	number of tokens	percent correct	Listener's response b	d	ɡ	p	t	k	ǰ	č
b	11	91	10			1				
d	15	47	3	7	1		4			
ɡ	10	60		1	6	1		2		
p	16	75	3			12	1			
t	14	93		1			13			
k	10	70				1	2	7		
ǰ	2	50		1					1	
č	2	50							1	1

(b) SE tokens

on 12 of the 94 tokens. The second candidate was correct in 9 of the 12 errors. In half of the errors, the system's top choice was listed as an alternate candidate by a spectrogram reader. A confusion matrix for the combined errors for all tasks is shown in Table 5.6. Averaged across all the tasks, 75% of the errors were in place of articulation and 17% were in voicing.

Most of the errors are reasonable, even though there may have been acoustic evidence for the correct answer. A few examples of tokens on which the system made errors are shown in Figure 5.6. The errors made on the two left tokens are more reasonable than the errors for the two tokens on the right. The leftmost token was called /d/, with /b/ as a second candidate. The burst release is located primarily at mid frequencies (2-4 kHz), a lower frequency location than is expected for an alveolar stop between front vowels. However, the preceding vowel is a fronted-/u/, which would probably not be fronted if

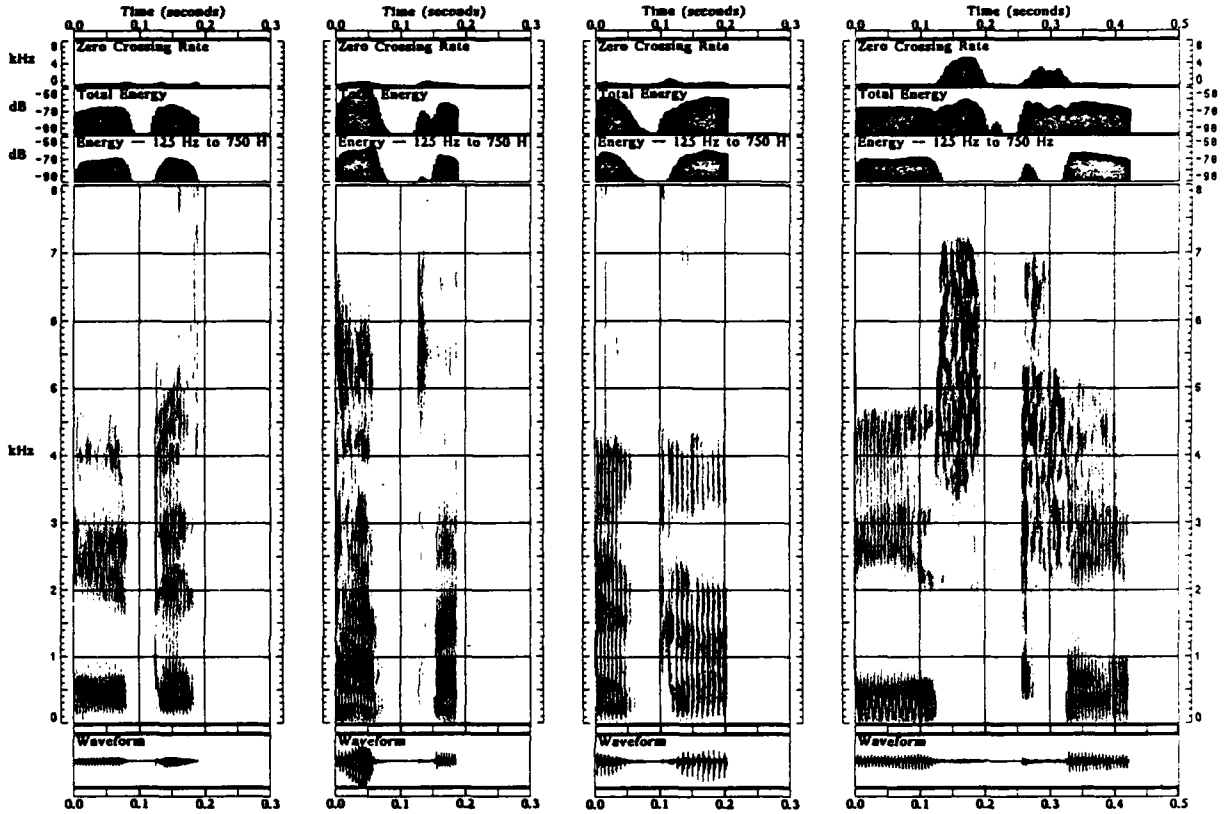
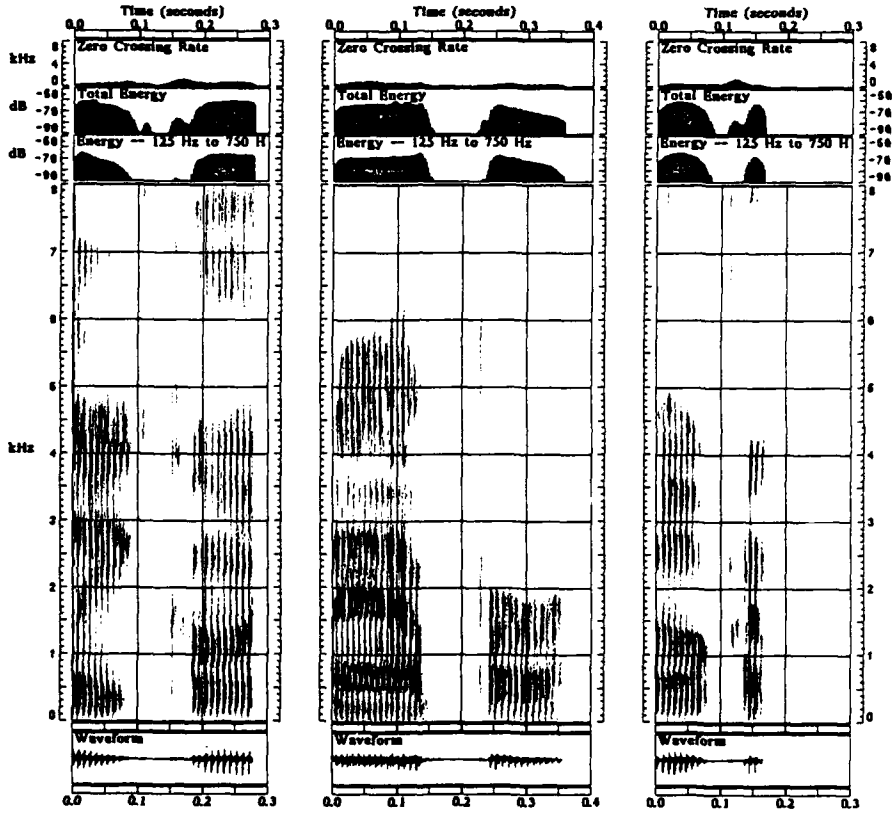


Figure 5.6: Examples of system errors on AC tokens.

the stop was labial. The system is not using the information that the preceding vowel is fronted. The formant transitions are also slightly better for alveolar than for labial.

The middle two spectrograms both have conflicting evidence between the burst characteristics and the formant transitions. In /ʌp–ə/ the formants are better for labial, but the high frequency concentration of energy in the release supports alveolar. While the weak release in /ɚ–ga/ suggests labial, the formant transitions from the stop into the /a/ are incompatible with labial. In this case, the system favored labial quite strongly, even though the spectrogram readers ruled labial out. While the rightmost spectrogram was called a /t/, /k/ was a close second choice. The distinguishing acoustic cue is subtle, but there is more energy in the release around 3 kHz than there is at higher frequencies, supporting velar.

Analysis of errors on the SE tokens: The tokens in the SE subset had an error by at least one spectrogram reader or one listener. The system made errors on 23 of the 80 tokens. The same error was made by a listener in 14 of these cases. A reader made the same error as the system on 11 of the tokens, and supplied the system's top choice as an alternate candidate for another 5. Only in 2 instances did the system propose an answer that did not agree with the listeners or the reader.



(a) /i–go/ and /æpɝ/ (b) /okə/

Figure 5.7: Examples of system errors on SE tokens.

A confusion matrix for the system's identification of the SE tokens is given in Table 5.6. The system made a higher proportion of voicing errors on the SE tokens than it did on the AC tokens. This is in agreement with both the listeners and the spectrogram readers on these tokens. Figure 5.7(a) shows spectrograms of two such errors. For the /ɡ/ on the left, the stop has a VOT that is neither short nor long, it is hard to tell if the stop is aspirated, and there is no prevoicing during the closure. The system made a voicing

error, calling the stop voiceless. The stop was correctly identified by the spectrogram reader as a /g/, with /k/ a second choice. Only 59% of the listeners correctly identified the stop as a /g/; the remainder called it /k/.

The /p/ in the middle has a short-VOT and no aspiration, and it may have a short period of prevoicing during closure. The prevoicing caused the system to favor a voiced stop over a voiceless one. The reader listed /p/ as the top choice and /b/ as an alternate. Nineteen of the 20 listeners identified the stop correctly; one listener called the stop /b/.

The spectrogram in (b) is an example of an error in the place of articulation that was also made by listeners and readers. The top candidate was a /p/, and /k/ was supplied as a second choice. The spectrogram reader labeled the stop only as /p/ as did 3 of the 20 listeners.

Performance with termination: The system was also evaluated using termination rules. If there was sufficient evidence for a phonetic feature, the system stopped attempting to determine that feature. Performance using the termination rules was essentially unchanged. An additional error occurred on task 1 and one fewer error was made on task 4. The termination rules applied 53% of the time for voicing, 48% of the time for place. and 27% of the time for both voicing and place. The system terminated early more often on AC tokens than on SE tokens.

Evaluation using other subjects to supply acoustic descriptions: The system has been evaluated on 24 tokens from the AC set with data supplied by users unfamiliar with the system. The system correctly identified 88% of the tokens, and proposed the correct answer as the second candidate when it made an error.

Evaluation on the SS-1 data: The rules used in the SS-1 system reported in Zue and Lamel (1986) were reimplemented in ART, with minimal modifications. The control structure in ART is primarily forward chaining, while in the MYCIN-based SS-1 system it was backward chaining. The reimplemented system had a "better feel" to it because of the difference in control strategy. In addition, the MYCIN-combine function was replaced by simple addition, and the certainties associated with the acoustic attributes in the SS-1 system were not used. Thus, a comparison has been made to determine what effect these design changes had on the system performance.

The ART implementation of the SS-1 rules (SS-2) was evaluated on the set 1 tokens used to evaluate the SS-1 system. The SS-2 system had an an accuracy of 88% for the top choice and 98% including the top two choices. The performance is comparable to that of SS-1 (see Table 5.1). This comparison indicates that the certainties associated with the acoustic attributes in the MYCIN-based system are not necessary. The change in the scoring strategy also seems to be unimportant.

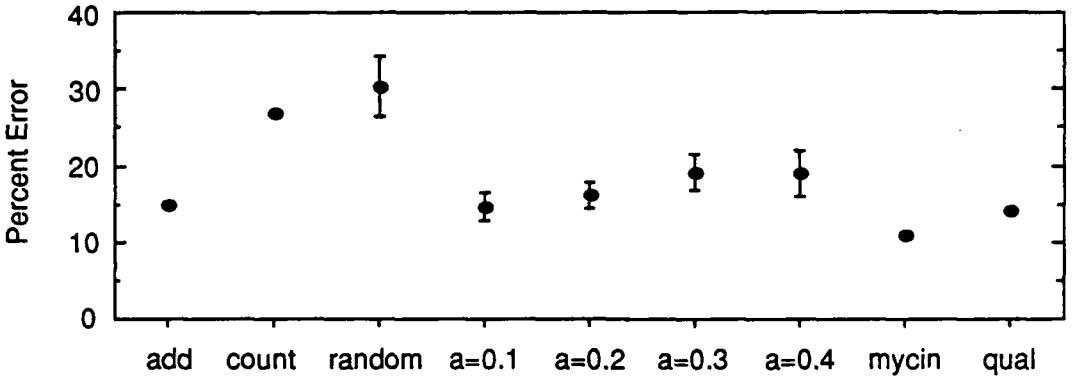


Figure 5.8: Comparison of scoring strategies on SS-1 set 1.

The same data were also used to assess the importance of the strengths associated with the rule conclusions. Two experiments were conducted. In the first experiment the rule strengths were all set to 1.0, eliminating the distinction between weak and strong evidence. The resulting error rate of 27%, shown as "count" in Figure 5.8, is almost double that of the baseline system which has a 15% error rate ("add" in Table 5.8). In a second experiment, a random number in the range [0, 1.0] for positive evidence and [-1.0, 0] for negative evidence, was generated and summed. This experiment was conducted 10 times. The mean error rate was 30% with a standard deviation of 4%. Both of these experiments indicate that rule strengths associated with the rule conclusions are important and that not all the evidence should be treated equally.

To evaluate the dependency of the performance on the selection of the numerical values, experiments were conducted in which a random number, in the range [-a,a], was added to the score at each update. The system was evaluated 10 times on set 1 for a = 0.1, 0.2, 0.3 and 0.4. The results are shown in Figure 5.8. The difference in the mean error rate from the baseline of 15% is insignificant at the .05 level for all cases. For a = 0.3, 0.4,

the difference is significant at the .01 level. These experiments indicate that the system is relatively insensitive to the numerical values assigned to the strengths.

The remaining two data points in Figure 5.8 show the performance of the system using two other scoring strategies discussed in section 5.6. The point labeled "mycin" used the EMYCIN-combine function (Shortliffe, 1976). The point labeled "qual" used the lexicographic scoring. That the system performance is comparable to the "add" method illustrates the robustness of the system to changes in scoring methods.

5.8 Discussion of some implementation issues

The hardest part of the system implementation, and by far the most time-consuming, was controlling the system to have acceptable behavior. Because of the way backward chaining is integrated in ART, different rules had to be written for almost every query in the system. Each query rule had a priority to ensure that it fired at the appropriate time.

The rules were structured so that "yes" evidence was used before "maybe" evidence. In order to simulate the concept "use information that is known to be true, before using evidence that might be true," the rules had to be duplicated, assigning a lower priority to rules based on uncertain evidence. It would have been better if a "meta-rule" could have been written that said to use certain evidence before using uncertain evidence. Duplicating rules requires extra care when the rules are modified, as the modifications may have to be made in multiple copies of the rule.

Another problem resulted from augmenting ART to handle confidences. The justification capabilities provided by ART could not be used. As a result justification rules to use for debugging and to provide explanations had to be written.

5.9 Summary

A description of the representation of phonemes and a set of rules relating phonemes to acoustic attributes has been provided. The reasoning of the system "feels" acceptable to a spectrogram reader. The performance of the system indicates that knowledge formalization has been somewhat successful. However, the ability of human spectrogram

readers and listeners surpasses that of the knowledge-based system, indicating the need for additional knowledge.

- The relationships between phonemes and acoustic attributes have been specified using a frame-like representation. An intermediate representation in terms of phonetic features was used.
- Rules were used to query the "user" for acoustic attributes. Phonetic features were deduced from qualitative acoustic attributes.
- The system simultaneously considered multiple hypotheses and maintained a ranking of hypotheses for each feature independently.
- The system could be run in a mode where it decided when it had enough evidence to believe the value of a feature, and stopped pursuing alternatives. The termination rules did not reduce the system's accuracy.
- Evaluation on a set of tokens from five tasks indicated that the errors made by the system were often reasonable and in agreement with spectrogram readers and listeners.
- The system was shown to be relatively insensitive to changes in scoring strategies. The experiments also indicated that the strengths associated with rule deductions were important.

Chapter 6

Concluding Remarks

Studying spectrograms of continuous speech provides an opportunity to learn about the canonical characteristics of speech sounds and coarticulatory effects (Fant, 1962; Klatt and Stevens, 1973). In order to obtain a better understanding of the spectrogram reading process, and to assess the abilities of spectrogram readers, spectrogram reading was formally evaluated. While a variety of spectrogram reading evaluations have been reported previously, it is difficult to assess and compare the experiments. The conditions of the experiments were quite varied, and only a few subjects were tested on data from a small number of talkers (typically 1-5). In Chapter 4 the results of rigorous testing of several spectrogram readers were discussed. The performance of five spectrogram readers, assessed on speech from almost 300 talkers, was found to be comparable to the best accuracies previously reported. The tasks at which the readers were assessed were quite difficult. The test tokens were extracted from continuous speech, spoken by many talkers, and in a variety of vowel and stress environments. Because the reader was provided with a spectrogram consisting of the one or two consonants to be identified and a single vowel on each side, lexical and other higher sources of knowledge could not be used. Spectrogram readers identified stop consonants with a high degree of accuracy despite the limited information available.

The performance of the spectrogram readers was compared to that of human listeners on the same tasks. In general, the listeners were able to identify stops more accurately than the readers could. Some of the difference may be due to experience; listeners have had much more experience listening than the readers have had at labeling spectrograms. The largest difference between listeners and readers was that while listeners correctly determined the place of articulation over 98% of the time, readers had about 90% accuracy

at place. This difference may be in part due to insufficiencies in the spectrographic representation. Spectrogram readers may sometimes have a difficult time resolving acoustic attributes, such as the direction and amount of formant motion, or the main location of energy in the stop release. Often there are several attributes for place in the spectrogram, which readers must selectively pay attention to. Errors by the readers typically occurred on tokens where ambiguities or conflicts in the acoustic characteristics were visible in the spectrogram.

Expert spectrogram readers have learned to interpret the patterns visible in the spectrogram and to form phonetic judgements. It is the knowledge that they have learned and the ability to assimilate it with an underlying knowledge of the principles of articulation, that contributes to their expertise. While the performance of listeners surpasses that of spectrogram readers, the information on which listeners base their decisions is not known. Spectrogram reading provides an opportunity to separate acoustic information from other sources, such as lexical, syntactic and semantic.

Knowledge obtained from spectrogram reading was incorporated in a rule-based system for stop identification. The emphasis was on capturing the acoustic descriptions and modeling the reasoning thought to be used by human spectrogram readers. Because there is ambiguity in relating acoustic events to the underlying phonemic representation, multiple descriptions and rules were used. As our knowledge about the acoustic correlates of speech sounds improves, the descriptions and rules can be modified.

Although this implementation does not reason from basic principles of speech production directly, these principles underly the knowledge representation and rule formulation. It would be interesting to explore a system that could reason from basic principles. Such a system would know how to relate articulatory information to the acoustic characteristics of speech sounds. For example, a principle of the system might be that the natural frequencies of an acoustic tube are inversely proportional to the length of the tube. Other properties of the system would account for the movements of the articulators and the different sources used in speech production.

My experience in implementing the knowledge-based system for stop identification has been that formalizing the knowledge used in spectrogram reading is much harder than I had anticipated. I believe that this is due to a variety of reasons. Perhaps the most important is that there appears to be much more happening in our visual system and in

our thought processes than we actually express, even when asked to explain our reasoning. An example is the human ability to selectively pay attention to acoustic evidence, particularly when there is contradictory evidence. In contrast, the system is relatively conservative and rarely ignores evidence, even in the presence of conflicting evidence. This is in part because there is no belief in the acoustic evidence supplied to the system. At times human experts will say things like "the formants are better for alveolar than velar, but I like the release so much better as velar, that I'm willing to ignore the formant transitions." The system is, by design, reluctant to use evidence as strongly as a spectrogram reader will, as the conditions under which readers do so are not well understood.

A related issue is the identification of the acoustic attributes. It may be relatively easy to develop algorithms to locate some of the attributes, such as location of the energy in the release, and the strength of the release. Other attributes, such as the presence of a double-burst or of aspiration, might be quite difficult. As another example, the formant transitions between stops and vowels may occur over a short time interval, such as the 30 ms before and after the stop. Humans are often able to determine the formant motion, while the problem formant tracking in general is still unsolved despite many efforts (McCandless, 1974; Shafer and Rabiner, 1969; Yegnanarayana, 1978; Talkin, 1987). In order to use automatically determined acoustic attributes it will be necessary to associate a goodness measure, or belief in the attribute. The problems of ambiguity and uncertainty (both in the acoustic attributes and the rules) are similar to those in diagnosis (Szolovitz and Pauker, 1978; Miller et al., 1984).

What about the applicability of rule-based systems to continuous speech recognition in general? There are several problems with this approach. Of primary importance is our level of understanding. Many more in-depth studies of continuous speech are needed before we can hope to have enough knowledge to build such a system. In addition, some of the problems in automatic speech recognition are probably handled better by signal processing or pattern matching techniques. For example, I did not develop rules to segment the acoustic signal. How, or whether, humans segment the spectrogram remains to be determined; readers may label the spectrogram without explicitly locating acoustic boundaries. Since acoustic segmentation is not well understood at this time, signal processing algorithms may be a more appropriate methodology (Glass, 1988). As such, building a complete continuous speech recognition system using only a rule-based

framework may not be the best application. However, using a rule-based system as a verification module, in limited tasks where our reasoning can be better quantified, may be advantageous for system development.

The primary reason for using a knowledge-based system was that the expression and use of the knowledge is explicit. Every decision made by the system is justified by supporting evidence. The system developer can inspect the rules and knowledge-base and understand the interactions. However, in this particular implementation, in order to cause the system to have reasonable behavior, some of the knowledge became obscured. For example, in order to implement concepts like "if the evidence is uncertain, loosen the constraints and believe the conclusions less," it was necessary to duplicate sets of rules. Rule duplication results in the same knowledge being expressed in a variety of places, making it harder to modify the system. Some systems use "meta"-rules to handle such cases (Davis, 1981).

This thesis has shown that some of the knowledge used in spectrogram reading can be expressed in acoustic attributes, which may be combined using context dependent rules to identify stops. The identification accuracy, across a wide variety of phonetic contexts and a large number of talkers, is about 5% below that of spectrogram readers. While these results are encouraging, there is still much more that can be learned from the study of speech. Spectrograms, and other time-frequency representations, are valuable aids for understanding the dynamic characteristics of continuous speech.

Bibliography

A. Ali, T. Gallagher, J. Goldstein, R. Daniloff (1971), "Perception of Coarticulated Nasality," *JASA*, vol. 49, no. 2, pp. 538-540.

J.B. Allen (1985), "Cochlear Modeling," *IEEE ASSP Magazine*, vol. 2, no. 1, pp. 3-29.

B.G. Buchanan and E.H. Shortliffe (1984), *Rule-based Expert Systems: The MYCIN experiments of the Heuristic Programming Project*, Reading, MA: Addison-Wesley.

M.A. Bush and G.E. Kopec (1987) "Network-Based Connected Digit Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 10, pp. 1401-1413.

M.A. Bush, G.E. Kopec, and V.W. Zue (1983) "Selecting Acoustic Features for Stop Consonant Identification," *Proc. IEEE ICASSP-83*, Boston, MA, pp. 742-725.

N. Carbonell, J.P. Haton, D. Fohr, F. Lonchamp, and J.M. Pierrel (1984), "An Expert System for the Automatic Reading of French Spectrograms," *Proc. IEEE ICASSP-84*, San Diego, CA, pp. 868-871.

N. Carbonell, J.P. Damestoy, D. Fohr, J.P. Haton, and F. Lonchamp (1986), "APHODEX, Design and Implementation of an Acoustic-Phonetic Decoding Expert System," *Proc. ICASSP-86*, Tokyo, Japan, pp. 1201-1204.

F.R. Chen (1985), *Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary*, Ph.D. Dissertation, Massachusetts Institute of Technology.

N. Chomsky and M. Halle (1968), *The Sound Pattern of English*, New York: Harper and Row.

W.J. Clancey (1984), "Classification Problem Solving," *Proc. National Conf. Artificial Intelligence, AAAI-84*, Austin, TX, pp. 49-55.

W.J. Clancey (1985), "Heuristic Classificaiton," *Artificial Intelligence*, vol. 27, no. 3, pp. 289-350.

W.J. Clancey and E.H. Shortliffe, eds. (1984), *Readings in Medical Artificial Intelligence: The First Decade*, Reading, MA: Addison-Wesley.

J.E. Clark (1983), "Intelligibility comparisons for two synthetic and one natural speech source," *J. Phonetics*, vol. 11, pp. 37-49.

J.R. Cohen (1986), "Application of an adaptive auditory model to speech recognition," *Proc. Symp. on Speech Recognition*, Montreal, pp. 8-9.

R.A. Cole, ed. (1980), *Perception and Production of Fluent Speech*, Hillsdale, NJ: Lawrence Erlbaum.

Bibliography

R.A. Cole, M.S. Phillips, S.M. Brill, P. Specker, and A.P. Pilant (1982), "Speaker-Independent Recognition of English Letters," *JASA*, vol. 72, Supp. 1, p. S31. (Paper presented at the 104th meeting of the Acoustical Society of America, Orlando, FL)

R.A. Cole, A.I. Rudnicky, and V.W. Zue (1979), "Performance of an Expert Spectrogram Reader," *JASA*, vol. 65, Supp. 1, p. S81. (Paper presented at the 97th meeting of the ASA, Boston, MA)

R.A. Cole, A.I. Rudnicky, V.W. Zue, and D.R. Reddy (1980), "Speech as Patterns on Paper," Chapter 1 in *Perception and Production of Fluent Speech*, R.A. Cole, ed., Hillsdale, NJ: Lawrence Erlbaum, pp. 3-50.

R.A. Cole, R.M. Stern, and M.J. Lasry (1985), "Performing Fine Phonetic Distinctions: Templates Versus Features," Chapter 15 in *Variability and Invariance in Speech Processes*, J.S. Perkell and D.H. Klatt, eds., Hillsdale, NJ: Lawrence Erlbaum, pp. 325-341.

R.A. Cole and V.W. Zue (1980), "Speech as Eyes See It," Chapter 2 in *Attention and Performance VIII*, R.S. Nickerson, ed. Hillsdale, NJ: Lawrence Erlbaum, pp. 475-494.

D.S. Cyphers (1985), *Spire: A Research Tool*, S.M. Thesis, Massachusetts Institute of Technology.

N.A. Daly (1987), *Recognition of Words from their Spellings: Integration of Multiple Knowledge Sources*, S.M. Thesis, Massachusetts Institute of Technology.

N. Davidsen-Nielson (1974), "Syllabification in English words with medial *sp, st, sk*," *J. Phonetics*, vol. 2, pp. 15-45.

R. Davis, "Interactive Transfer of Expertise: Acquisition of New Inference Rules," in *Readings in Artificial Intelligence*, B.L. Webber and N.J. Nilsson, eds., Palo Alto, CA: Tioga, pp. 410-428.

R. Davis and D.B. Lenat, eds. (1982), *Knowledge-Based Systems in Artificial Intelligence*, New York: McGraw-Hill.

R. Davis, B. Buchanan, and E. Shortliffe (1977), "Production Rules as a Representation for a Knowledge-Based Consultation Program," *Artificial Intelligence*, vol. 8, no. 1, pp. 15-45.

P.C. Delattre (1951), "The Physiological Interpretation of Sound Spectrograms," *PMLA*, vol. LXVI, no. 5, pp. 864-875.

P.C. Delattre, A.M. Liberman, and F.S. Cooper (1955), "Acoustic Loci and Transitional Cues for Consonants," *JASA*, vol. 27, no. 4, pp. 769-773.

P. Demichelis, R. De Mori, P. Laface, and M. O'Kane (1983), "Computer Recognition of Plosive Sounds Using Contextual Information," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-31, no. 2, pp. 359-377.

R. De Mori, P. Laface, and Y. Mong (1985), "Parallel Algorithms for Syllable Recognition in Continuous Speech," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-7, no. 1, pp. 56-69.

R. De Mori (1983), *Computer Models of Speech Using Fuzzy Algorithms*,New York: Plenum.

P.B. Denes (1955), "Effect of Duration in the Perception of Voicing," *JASA*, vol. 27, no. 4, pp. 761-764.

P.B. Denes (1963), "On the Statistics of Spoken English," *JASA*, vol. 35, no. 6, pp. 892-904.

R.O. Duda and J.G. Gaschnig (1981), "Knowledge-Based Expert Systems Come of Age," *Byte*, p. 238.

R. Duda, J. Gaschnig, and P. Hart (1981), "Model Design in the PROSPECTOR Consultant System for Mineral Exploration," in *Readings in Artificial Intelligence*, B.L. Webber and N.J. Nilsson, eds., Palo Alto, CA: Tioga, pp. 192-199.

R.O. Duda, P.E. Hart, and N.J. Nilsson (1976), "Subjective Bayesian methods for rule-based inference systems," Proc. National Computer Conf., AFIPS Conf. Proc. vol. 45, pp. 1075-1082.

T. Edwards (1981), "Multiple feature analysis of intervocalic English plosives", *JASA*, vol. 69, no. 2, pp. 535-547.

J. Egan (1944), "Articulation testing methods II," OSRD Report No. 3802, U.S. Dept. of Commerce Report PB 22848.

L.D. Erman and V.R. Lesser (1980), "The Hearsay-II Speech Understanding System: A Tutorial," Chapter 16 in *Trends in Speech Recognition*, W.A. Lea, ed., Englewood Cliffs, NJ: Prentice-Hall, pp. 361-181.

C. Espy-Wilson (1987), *An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels*, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

G. Fairbanks (1958), "Test of Phonemic Differentiation: The Rhyme Test," *JASA*, vol. 30, no. 7, pp. 596-600.

G. Fant (1960), *Acoustic Theory of Speech Production*, The Hague: Mouton.

G. Fant (1962), "Descriptive Analysis of the Acoustic Aspects of Speech," *Logos*, vol. 5, no. 1, pp. 3-17.

G. Fant (1968), "Analysis and synthesis of speech process" Chapter 8 in *Manual of Phonetics*, B. Malmberg, ed., Amsterdam, The Netherlands: North-Holland, pp. 173-277.

G. Fant (1973), *Speech Sounds and Features*, Cambridge, MA: MIT Press.

E. Fischer-Jørgensen (1954), "Acoustic Analysis of Stop Consonants," *Miscellenea Phonetica*, vol. II, pp. 42-59. (also Chapter 11 in *Readings in Acoustic Phonetics*, I. Lehiste, ed. (1967), Cambridge, MA: MIT Press)

W. Fisher, G. Doddington, and K. Goudie-Marshall (1986), "The DARPA Speech Recognition Research Database: Specifications and Status," Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, pp. 93-99.

J.L. Flanagan (1972), *Speech Analysis Synthesis and Perception*, Berlin: Springer-Verlag.

O. Fujimura (1960), "Spectra of Nasalized Vowels," *Research Laboratory of Electronics, MIT Quarterly Report*, no. 58, pp. 214-218.

O. Fujimura (1962), "Analysis of Nasal Consonants," *JASA*, vol. 34, no. 12, pp. 1865-1875.

J.G. Gaschnig (1982), "Prospector: An Expert System For Mineral Exploration," Chapter 3 in *Introductory Readings in Expert Systems*, D. Michie, ed., New York: Gordon and Breach, pp. 47-64.

O. Ghitza (1987), "Robustness against noise: The role of timing-synchrony measurement," *Proc. ICASSP-87*, Dallas, TX, pp. 2372-2375.

O. Ghitza (1988), "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. Phonetics*, vol. 16, no. 1, pp. 109-123.

J.R. Glass (1984), *Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment*, S.M. Thesis, Massachusetts Institute of Technology.

J.R. Glass and V.W. Zue (1988), "Multi-level Acoustic Segmentation of Continuous Speech," *Proc. IEEE ICASSP-88*, New York, NY, pp. 429-433.

R. Goldhor (1985), *Representation of consonants in the peripheral auditory system: A modeling study of the correspondence between response properties and phonetic features*, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

B.G. Greene, D.B. Pisoni, and T.D. Carrell (1984), "Recognition of speech spectrograms," *JASA*, vol. 76, no. 1, pp. 32-43.

W.E.L. Grimson and R.S. Patil, eds. (1987), *AI in the 1980s and Beyond*, Cambridge, MA: MIT Press.

F. Grosjean and J.P. Gee (1984), "Another View of Spoken Word Recognition," Working Paper, Northeastern University, Boston, MA.

M. Halle, G.W. Hughes and J.-P.A. Radley (1957), "Acoustic Properties of Stop Consonants," *JASA*, vol. 29, no. 1, pp. 107-116.

J.P. Haton and J.P. Damestoy (1985), "A Frame Language for the Control of Phonetic Decoding in Continuous Speech Recognition," *Proc. IEEE ICASSP-85*, Tampa, FL, pp. 41.6.1-4.

F. Hayes-Roth (1984), "The Knowledge-Based Expert System: A Tutorial," *Computer*, vol. 17, no. 9, pp. 11-28.

F. Hayes-Roth, D.A. Waterman, and D.B. Lenat (1983), *Building Expert Systems*, London: Addison-Wesley.

J.T. Hogan and A.J. Rozsypal (1979), "Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant," *JASA*, vol. 65, no. 5, pp. 1764-1771.

R.A. Houde and J.L Braeges (1983) "Independent Drill: A role for speech training aids in the speech development of the deaf," Chapter 16 in I. Hockberg, H. Levitt, and M.J. Osberger, eds., *Speech of the Hearing Impaired: Research, training, and personnel preparation*, Baltimore, MD: University Park Press.

A.S. House and G. Fairbanks (1953), "The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels," *JASA*, vol. 25, no. 1, pp. 105-113.

A.S. House, D.P. Goldstein, and G.W. Hughes (1968), "Perception of Visual Transforms of Speech Stimuli: Learning Simple Syllables," *American Annals of the Deaf*, vol. 113, no. 2, pp. 215-221.

A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter (1965) "Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set," *JASA*, vol. 37, no. 1, pp. 158-166.

L.S. Hultzen (1965), "Consonant Clusters in English," *American Speech*, vol. XL, no. 1, pp. 5-19.

R. Jacobson, C.G.M. Fant, and M. Halle (1952), "Preliminaries to Speech Analysis," MIT Acoustics Laboratory, Technical Report No. 13.

F. Jelinek (1976), "Continuous Speech Recognition by Statistical Methods," *Proc. IEEE,* vol. 64, no. 4, pp. 532-556.

J. Johannsen, J. MacAllister, T. Michalek, and S. Ross (1983), "A Speech Spectrogram Expert," *Proc. IEEE ICASSP-83*, Boston, MA, pp. 746-749.

S.R. Johnson, J.H. Connolly, and E.A. Edmonds (1984), "Spectrogram Analysis: A Knowledge-Based Approach to Automatic Speech Recognition," Leicester Polytechnic, Human Computer Interface Research Unit, Report No. 1.

D. Kahn (1976), *Syllable-based Generalizations in English Phonology*, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

L.N. Kanal and J.F. Lemmer, eds. (1986), *Uncertainty in Artificial Intelligence*, New York: North Holland.

P.A. Keating, J.R. Westbury, and K.N. Stevens (1980), "Mechanisms of stop-consonant release for different places of articulation," *JASA*, Supp. 1, vol. 67, p. S93. (Paper presented at the 99th meeting of the ASA, Atlanta, GA)

R.L. Keeney and H. Raiffa (1976), *Decisions with Multiple Variables: Preferences and Value Tradeoffs*, New York: John Wiley & Sons.

D.H. Klatt (1975), "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," *JSHR*, vol. 18, no. 4, pp. 686-706.

D.H. Klatt (1976), "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *JASA*, vol. 59, no. 5, pp. 1208-1221.

D.H. Klatt (1977), "Review of the ARPA Speech Understanding Project," *JASA*, vol. 62, no. 6, pp. 1345-1366.

D.H. Klatt and K.N. Stevens (1973), "On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram-Reading Experiment," *IEEE Trans. Audio and Electroacoustics*, vol. AU-21, no. 3, pp. 210-217.

W. Koenig, H.K. Dunn, and L.Y. Lacey (1946), "The Sound Spectrograph," *JASA*, vol. 18, no. 1, pp. 19-49.

G. Kopec (1984), "Voiceless stop consonant identification using LPC spectra", *Proc. IEEE ICASSP-84*, San Diego, CA., pp. 42.1.1-4.

G.M. Kuhn and R.McI. McGuire (1974), "Results of a VCV Spectrogram-Reading Experiment," Haskins Laboratories, Status Report on Speech Research SR-39/40, pp. 67-80.

B. Kuipers and J.P. Kassirer (1984), "Causal Reasoning in Medicine: Analysis of Protocol," *Cognitive Science*, vol. 8, pp. 363-385.

L.F. Lamel (1985), "Stop Identification from Spectrograms," Term Project Report for MIT Course 6.871, Spring 1985.

L.F. Lamel, R.H. Kassel, and S. Seneff (1986), "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, pp. 100-109.

Bibliography

L.F. Lamel and V.W. Zue (1984), "Properties of Consonant Sequences within Words and Across Word Boundaries," *Proc. IEEE ICASSP-84*, San Diego, CA, pp. 42.3.1-4.

W.A. Lea, ed. (1980), *Trends in Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall.

I. Lehiste (1962), "Acoustical Characteristics of Selected English Consonants," Report No. 9, U. Michigan, Communication Sciences Laboratory, Ann Arbor, Michigan.

I. Lehiste, ed. (1967), *Readings in Acoustic Phonetics*, Cambridge, MA: MIT Press.

V.R. Lesser, R.D. Fennell, L.D. Erman, and D.R. Reddy (1975), "Organization of the Hearsay II Speech Understanding System," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 1, pp. 11-24.

A.M. Liberman, F.S. Cooper, D.P. Shankweiler, M. Studdert-Kennedy (1967), "Perception of the Speech Code," *Psychological Review*, vol. 74, pp. 431-461.

A.M. Liberman, F.S. Cooper, D.P. Shankweiler, M. Studdert-Kennedy (1968), "Why are Speech Spectrograms Hard to Read," *American Annals of the Deaf*, vol. 113, pp. 127-133.

B.J. Lindblom and S. Svensson (1973), "Interaction Between Segmental and Nonsegmental Factors in Speech Recognition," *IEEE Trans. Audio and Electroacoustics*, vol. AU-21, no. 6, pp. 536-545.

L. Lisker (1957), "Closure Duration and the Intervocalic Voiced-Voiceless Distinction in English," *Language*, vol. 33, pp. 42-49.

L. Lisker (1978), "Rapid vs. Rabid: A Catalogue of Acoustic Features That May Cue the Distinction," Haskins Laboratories: Status Report on Speech Research SR-54, pp. 127-132.

L. Lisker and A.S. Abramson (1964), "A cross-Language Study of Voicing in Initial Stops: Acoustic Measurements," *Word*, vol. 20, no. 3, pp. 384-422.

F. Lonchamp (1985), "Reading Spectrograms: The View from the Expert," Preliminary version of a tutorial given at *Principes de la communication Homme-Machine: parole, vision et langage nature* Versilles, France.

B. Lowerre and D.R. Reddy (1980), "The Harpy Speech Understanding System," Chapter 15 in *Trends in Speech Recognition*, W.A. Lea, ed., Englewood Cliffs, NJ: Prentice-Hall, pp. 340-360.

J M. Lucassen (1985), "Building a Spectrogram Reader's Assistant as a Feasibility Study for a Fully Automated Spectrogram Reader," Term Project Report for MIT Course 6.871, Spring 1985.

R.F. Lyon (1984), "Computational Models of Neural Auditory Processing," *Proc. IEEE ICASSP-84*, San Diego, CA, pp. 36.1.1-4.

A. Malécot (1960), "Vowel Nasality as a Distinctive Feature in American English," *Language*, vol. 36, no. 2, pp. 222-229.

D. Marr (1982), *Vision*, San Francisco, CA: W.H. Freeman.

S. McCandless (1974), "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-22, no. 2, pp. 135-141.

J. McDermott (1982), "R1: A Rule-Based Configurer of Computer Systems," *Artificial Intelligence*, vol. 19, no. 1, pp. 39-88.

M.F. Medress (1980), "The Sperry Univac System for Continuous Speech Recognition," Chapter 19 in *Trends in Speech Recognition*, W.A. Lea, ed., Englewood Cliffs, NJ: Prentice-Hall, pp. 445-460.

P. Mermelstein (1977), "On Detecting Nasals in Continuous Speech," *JASA*, vol. 61, no. 2, pp. 581-587.

D. Michie, ed. (1982), *Introductory Readings in Expert Systems*, New York: Gordon and Breach.

G.A. Miller and P.E. Nicely (1955), "An Analysis of Perceptual Confusions Among Some English Consonants," *JASA*, vol. 27, no. 2, pp. 338-352.

R.A. Miller, H.E. Pople, and J.D Myers (1984), "INTERNIST-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine," Chapter 8 in *Readings in Medical Artificial Intelligence*, W. Clancey and E. Shortliffe, eds., Reading, MA: Addison-Wesley, pp. 190-209.

M. Minsky (1975), "A Framework for Representing Knowledge," Chapter 6 in *The Psychology of Computer Vision*, P.H. Winston, ed., New York: McGraw-Hill, pp. 211-277.

L. Molho (1975), "Automatic acoustic analysis of fricative and plosives", *Proc. ICASSP-76*, Philadelphia, PA, pp. 182-185.

A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J.C.R. Licklider, J. Munson, R. Reddy, and W. Woods (1971), *Speech Understanding Systems: Final Report of a Study Group*, (Reprinted 1973 Amsterdam, Netherlands: North-Holland/American Elsevier).

R.S. Nickerson (1978), "On the role of vision in language acquisition by deaf children," Chapter 7 in *Deaf Children: Developmental Perspectives*, New York: Academic Press.

R.S. Nickerson and A.W.F. Huggins (1977), "Assessment of Speech Quality," Bolt Beranek and Newman, Inc., Report No. 3486, Cambridge, MA.

R.S. Nickerson and K.N. Stevens (1973), "Teaching Speech to the Deaf: Can a Computer Help," *IEEE Trans. Audio and Electroacoustics*, vol. AU-21, no. 5.

H.C. Nusbaum, M.J. Dedina, and D.B. Pisoni (1984), "Perceptual Confusions of Consonants in Natural and Synthetic CV Syllables," Research on Speech Perception Progress Report No. 10, Indiana University, pp. 409-422.

P.W. Nye and J.H. Gaitenby (1973), "Consonant Intelligibility in Synthetic Speech and in Natural Speech Control (Modified Rhyme Test Results)," Haskins Laboratories, Status Report on Speech Research SR-33, pp. 77-91.

B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu, and J. Aurbach (1975), "The Role of Phonological Rules in Speech Understanding Research," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-23, pp. 104-112.

S.G. Pauker, G.A. Gorry, J.P. Kassirer, and W.B. Schwartz (1976), "Towards the Simulation of Clinical Cognition Taking a Present Illness by Computer," *American J. Medicine*, vol. 60, pp. 981-996.

G.E. Peterson and I. Lehiste (1960), "Duration of syllable nuclei in English," *JASA*, vol. 32, no. 1, pp. 693-703. (also Chapter 15 in *Readings in Acoustic Phonetics*, I. Lehiste, ed. (1967), Cambridge, MA: MIT Press)

Bibliography

M.S. Phillips (1987), "Speaker Independent Classification of Vowels and Diphthongs in Continuous Speech," *Proc. Eleventh International Congress of Phonetic Sciences*, Tallinn, Estonia, U.S.S.R., vol. 5, Se 85.3, pp. 240-243.

J.M. Pickett and I. Pollack (1964), "Intelligibility of Excerpts from Fluent Speech: Effects of Rate of Utterance and Duration of Excerpt," *Language and Speech*, vol. 6, part 3, pp. 151-164.

D.B. Pisoni and S. Hunnicutt (1980), "Perceptual Evaluation of MITALK: The MIT Unrestricted Text-to-Speech System," *Proc. ICASSP-80*, Denver, CO, pp. 572-275.

D.B. Pisoni, B.G. Greene, and T.D. Carrell (1983), "Identification of Visual Dispays of Speech: Comparisons of Naive and Trained Observers," *JASA*, vol. 73, supp. 1, p. S102. (Paper presented at the 105th meeting of the Acoustical Society of America, Cincinnati, OH)

R.K. Potter, G.A. Kopp and H.G. Kopp (1966), *Visible Speech*, New York: Dover.

M.A. Randolph (1985), "The application of a hierarchical classification technique to speech analysis," *JASA*, Supp. 1, vol. 77, p. S10. (Paper presented at the 109th meeting of the ASA, Austin, TX)

L.J. Raphael, M.F. Dorman, and F. Freeman (1975), "Vowel and Nasal Duration as Cues to Voicing in Word-Final Stop Consonants: Spectrographic and Perceptual Studies," *JSHR*, vol. 18, no. 3, pp. 389-400.

D.R. Reddy (1976), "Speech Recognition by Machine: A Review," *Proc. IEEE*, vol. 64, no. 4, pp. 501-531.

M. Rothenberg (1963), "Programmed Learning Problem Set, to teach the interpretation of a class of speech spectrograms," Ann Arbor, MI: Ann Arbor.

S. Roucos and M.O. Dunham (1987), "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *Proc. IEEE ICASSP-87*, Dallas, TX, pp. 73-77.

C.L. Searle, J.J. Zachary, and S.G. Rayment (1979), "Stop Consonant Discrimination Based on Human Audition," *JASA*, vol. 65, no. 3, pp. 799-809.

C.L. Searle, J.Z. Jacobson, and B.P. Kimberley (1980), "Speech as patterns in the 3-space of time and frequency," Chapter 3 in *Perception and Production of Fluent Speech*, R.A. Cole, ed., Hillsdale, NJ: Lawrence Erlbaum, pp. 73-102.

S. Seneff (1979), "A Spectrogram Reading Experiment," Term paper for MIT Course on Sound, Speech, and Hearing.

S. Seneff (1988), "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, no. 1, pp. 55-76.

S. Shamma (1988), "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives," *J. Phonetics*, vol. 16, no. 1, pp. 77-91.

D.W. Shipman (1982), "Development of speech research software on the MIT lisp machine," *JASA*, Supp. 1, vol. 71, p. S103. (Paper presented at the 103rd meeting of the ASA, Chicago, IL)

L. Shockey and D.R. Reddy (1975), "Quantitative Analysis of Speech Perception," in *Proceedings of the Stockholm Speech Communication Seminar*, G. Fant, ed., New York: John Wiley and Sons.

R. Shafer, L. Rabiner (1969), "System for Automatic Formant Analysis of Voiced Speech," *JASA*, vol. 47, no. 2, pp. 634-648.

E.H. Shortliffe (1976), *Computer Based Medical Consultations: MYCIN*, New York: American Elsevier.

A.I. Solzhenitsyn (1968), *The First Circle*, Translated from Russian by T.P. Whitney, New York: Harper & Row.

P.-E. Stern (1986), *Un Systeme Expert en Lecture de Spectrogrammes*, Ph.D. Dissertation, Universite de Paris-Sud, Centre D'Orsay.

P.-E. Stern, M. Eskenazi, and D. Memmi (1986), "An expert system for speech spectrogram reading," *Proc. IEEE ICASSP-86*, Tokyo, Japan, pp. 1193-1196.

K.N. Stevens (1980), "Acoustic correlates of some phonetic categories," *JASA*, vol. 63, no. 3, pp. 836-842.

K.N. Stevens, "Basic Acoustics of Vocal-Tract Resonators," Chapter 4 of a book on Acoustic Phonetics, in preparation.

K.N. Stevens and S.E. Blumstein (1981), "The search for invariant acoustic correlates of phonetic features," Chapter 1 in *Perspectives on the study of speech,*, P.D. Eimas and J.L. Miller, eds., Hillsdale, NJ: Lawrence Erlbaum, pp. 2-38.

K.N. Stevens and D.H. Klatt (1974), "Role of formant transitions in the voiced-voiceless distinction for stops," *JASA*, vol. 55, no. 3, pp. 653-659.

L.C. Stewart, R.A. Houde, and W.D. Larkin (1976), "A Real Time Sound Spectrograph with Implications for Speech Training for the Deaf," *Proc. IEEE ICASSP-76*, Philadelphia, PA, pp. 590-593.

S.G. Svenssen, (1974) *Prosody and Grammar in Speech Perception*, Monographs from the Institute of Linguistics, University of Stockholm, (MILOS), vol. 2.

P. Szolovitz, ed. (1982), *Artificial Intelligence in Medicine*, vol. 451 of AAAS Selected Symposium Series, Boulder, CO: Westview Press.

P. Szolovitz and S.G. Pauker (1978), "Categorical and Probabilistic Reasoning in Medical Diagnosis," *Artificial Intelligence*, vol. 11, pp. 115-144.

D. Talkin (1987), "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *JASA*, Supp. 1, vol. 82, p. S55. (Paper presented at the 114th meeting of the ASA, Miami, FL)

K. Tanaka (1981), "A parametric representation and clustering method for phoneme recognition - application to stops in a CV environment", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 6, pp. 1117-1127.

N. Umeda (1975), "Vowel Duration in American English," *JASA*, vol. 58, no. 2, pp. 434-445.

N. Umeda (1977), "Consonant Duration in American English," *JASA*, vol. 61, no. 3, pp. 846-858.

J. Vaissiere (1983), "Speech Recognition: A tutorial," course given in Cambridge, UK, July, 1983.

Bibliography

W.S-Y. Wang, and J. Crawford (1960), "Frequency Studies of English Consonants," *Language and Speech*, vol. 3, pp. 131-139.

D.A. Waterman (1986), *A Guide to Expert Systems*, Reading, MA: Addison-Wesley.

D.A. Waterman and F. Hayes-Roth, eds. (1978), *Pattern-directed inference systems*, New York: Academic Press.

M. Webster (1964), *Pocket Dictionary*, computer readable form.

C.J. Weinstein, S.S. McCandless, L.F. Mondstein, and V.W. Zue, (1975), "A System for Acoustic-Phonetic Analysis of Continuous Speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 1, pp. 54-67.

P.H. Winston (1984), *Artificial Intelligence*, Reading, MA: Addison-Wesley.

P.H. Winston and K.A. Prendergast, eds. (1985), *The AI Business: The commercial uses of Arificial Intelligence*, Cambridge, MA: MIT Press.

H. Winitz, M. Scheib, and J. Reeds (1972), "Identification of Stops and Vowels for the Burst Portions of /p,t,k/ Isolated from Conversational Speech," *JASA*, vol. 51, no. 4, pt. 2, pp. 1309-1317.

W. Woods, M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, and V. Zue (1976), "Speech Understanding Systems: Final Technical Progress Report," Bolt Beranek and Newman, Inc., Report No. 3438, Cambridge, MA.

B. Yegnanarayana (1978), "Formant Extraction from Linear-Prediction Phase Spectra," *JASA*, vol. 63, no. 5, pp. 1638-1640.

L.A. Zadeh, K.-S. Fu, K. Tanaka, and M. Shimura, eds. (1975), *Fuzzy Sets and their Applications to Cognitive and Decision Processes*, New York: Academic.

V.W. Zue (1976), "Acoustic Characteristics of Stop Consonants: A Controlled Study," Ph.D. Dissertation, Massachusetts Institute of Technology. (also reprinted by IULC, 1980)

V.W. Zue (1981), "Acoutic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments," presented at the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France, 1981.

V.W. Zue (1985), "Notes on Spectrogram Reading," Course notes for MIT Special Summer Course Speech Spectrogram Reading: An Acoustic Study of English Words and Sentences, Cambridge, MA.

V.W. Zue (1985a), "Acoustic Theory of Speech Production," Supplementary notes for MIT Course 6.343 Digital Speech Processing, Cambridge, MA.

V.W. Zue (1986), "The Use of Speech Knowledge in Automatic Speech Recognition," *Proc. IEEE Special Issue on man-machine speech communication*, vol. 73, no. 11, pp. 1602-1615.

V.W. Zue and R.A. Cole (1979), "Experiments on Spectrogram Reading," *Proc. IEEE ICASSP-79*, Washington, D.C., pp. 116-119.

V.W. Zue, D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M. Randolph, S. Seneff, J.E. Unverferth,III, and T. Wilson (1986), "The Development of the MIT Lisp-Machine Based Speech Research Workstation," *Proc. IEEE ICASSP-86*, Tokyo, Japan, pp. 329-332.

V.W. Zue and M. Laferriere (1979), "Acoustic Study of Medial /t,d/ in American English," *JASA*, vol. 66, no. 4, pp. 1039-1050.

V.W. Zue and L.F. Lamel (1986), "An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition." *Proc. ICASSP-86*, Tokyo, Japan, pp. 23.2.1-4.

V.W. Zue and E.B. Sia (1982), "Nasal Articulation in Homorganic Clusters in American English," Speech Communication Group Working Papers, Massachusetts Institute of Technology, vol. 1, pp. 9-18.

Appendix A

Spectrogram reading token sets

Table A.1 shows the percent of overall listener identification rate and the percent of tokens that were heard correctly (AC) by all listeners for each task. In selecting the *balanced set*, *B*, tokens for the readers, I tried to obtain the same distributions. The resulting distributions for the spectrogram test sets are given in Table A.1. Since there were multiple test sets for tasks 1, 2, and 4, both the individual and summed data are provided.

Table A.1: Error statistics for (top) listening and (bottom) reading tasks.

Task	Number of tokens	Identification rate	Percent AC (all correct)
1	633	97	78
2	313	88	59
3	312	96	81
4	175	85	48
5	160	93	73

Task	Number of tokens	Percent All Correct (AC)		
		Total	Balanced set	Extra tokens
1-1	52	42	79	0
1-2	53	42	79	0
1-3	52	42	79	0
1-4	53	42	79	0
1	210	42	79	0
2-1	52	42	59	0
2-2	50	42	60	0
2	102	42	60	0
3	51	45	77	12
4-1	51	37	50	24
4-2	51	33	48	19
4-3	51	35	48	23
4	153	35	49	22
5	46	59	73	36

Appendix A. Spectrogram reading token sets

Table A.2 provides information for the spectrogram test sets with regard to the number of talkers, sex of the talkers, and the sentence corpus from which the tokens were extracted. The number of different vowel contexts are also given.

Table A.2: Distribution of tokens for task test sets: (top) sex and database, and (bottom) vowel context.

Task	Number of tokens	Percent TIMIT	Percent IC	Number of talkers	Percent male	Percent female
1-1	52	58	42	50	52	48
1-2	53	57	43	50	51	49
1-3	52	56	44	47	48	52
1-4	53	57	43	50	55	45
1	210	57	43	164	51	49
2-1	52	62	38	45	46	54
2-2	50	62	38	49	54	46
2	102	62	38	85	54	46
3	51	59	41	46	41	59
4-1	51	57	43	51	57	43
4-2	51	59	41	46	49	51
4-3	51	63	37	50	51	49
4	153	59	41	128	48	52
5	46	67	33	43	59	41

Task	Number of tokens	Number of preceding vowels	Number of following vowels	Number of vowel contexts
1-1	52	13	15	41
1-2	53	12	16	47
1-3	52	12	17	44
1-4	53	12	18	48
1	210	14	18	101
2-1	52	13	15	39
2-2	50	12	17	43
2	102	13	17	60
3	51	10	15	32
4-1	51	16	11	40
4-2	51	16	11	46
4-3	51	17	12	43
4	153	17	13	87
5	46	13	12	36

Appendix B

Listeners' identification of tokens in spectrogram sets

Tables B.1—B.5 are confusion matrices for the listeners on the subset of tokens used in the spectrogram reading tests. While the tables are similar to those presented in Chapter 4, these provide a more accurate comparison to the readers' results.

Table B.1: Confusion matrix for listeners' identification of syllable-initial singleton stops in spectrogram sets.

Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	42	91.5	1114	22	4	73	1		4
d	32	89.4	9	830	22	1	58		8
g	38	85.3	4	3	940	1		152	2
p	31	93.6	46			841	5	1	6
t	34	94.4	1	36	9	3	931	3	3
k	33	95.2	6		31	4	2	911	3

Table B.2: Confusion matrix for listeners' identification of syllable-initial stops preceded by alveolar strong fricatives in spectrogram sets.

Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	15	77.8	105			27	2	1	
d	16	59.0		85	3		56		
g	8	76.4			55			17	
p	18	89.5	17			145			
t	24	87.9		23		2	190		1
k	21	92.6			14			175	

Table B.3: Confusion matrix for listeners' identification of syllable-initial stop-semivowel clusters and syllable-initial affricates in spectrogram test set.

Answer	Number of tokens	Percent correct	Listener's response								
			b	d	g	p	t	k	ǰ	č	none
b	7	91.4	64		2	4					
d	6	65.0		39					21		
g	3	100.0			30						
p	4	87.5	1			35	2				2
t	9	91.1	1				82	1		5	1
k	7	98.6						69			1
ǰ	7	71.4		2			4		50	14	
č	8	90.0					6		2	72	

Table B.4: Confusion matrix for listeners' identification of non-syllable-initial singleton stops in spectrogram sets.

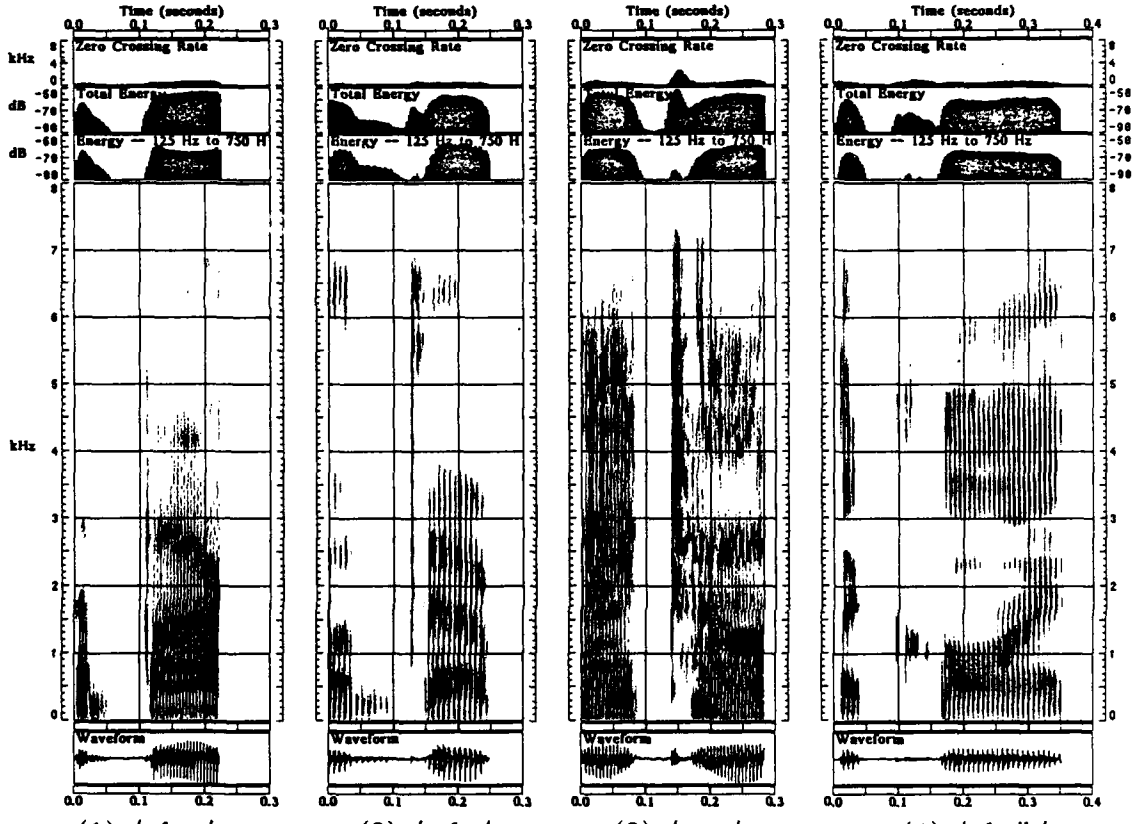
Answer	Number of tokens	Percent correct	*Listener's response*						
			b	d	g	p	t	k	none
b	12	92.9	223	4	1	7	2	1	2
d	50	93.8	2	469	2		27		
g	11	91.4			201			19	
p	36	85.1	87	1	3	613	8	8	
t	35	58.4	3	278	4	2	409	3	1
k	34	93.4	1	1	22	5	15	635	1

Table B.5: Confusion matrix for listeners' identification of non-syllable-initial stops in homorganic nasal-stop clusters in spectrogram set.

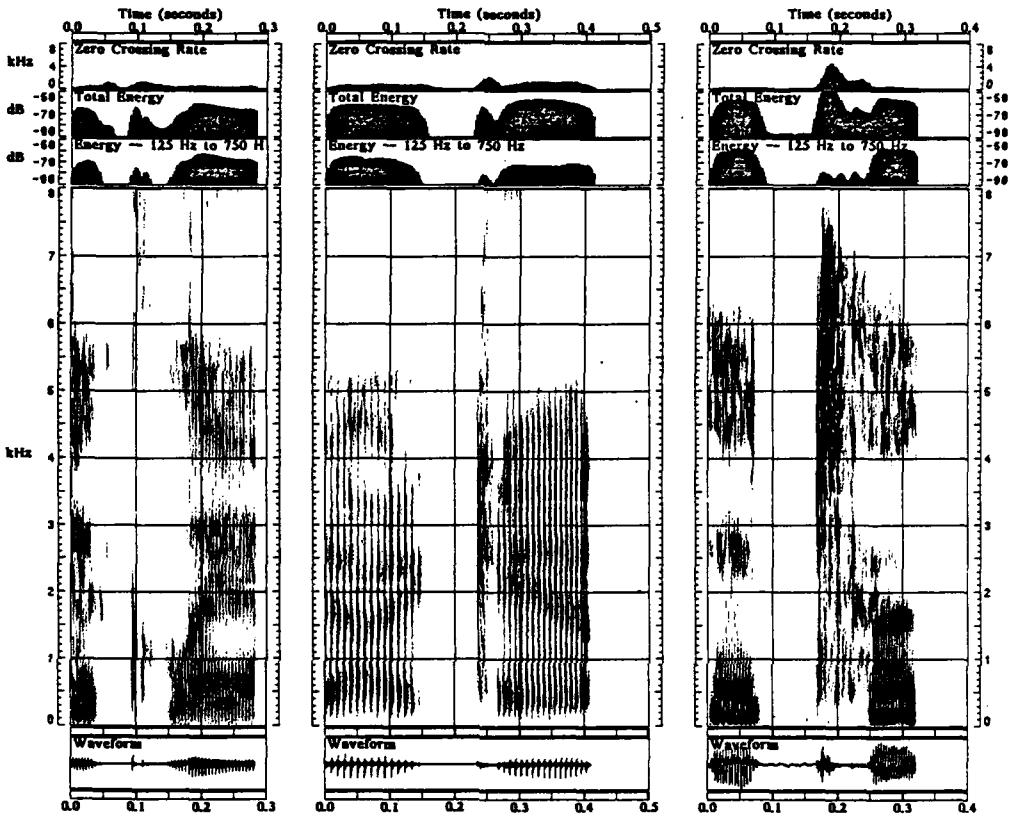
Answer	Number of tokens	Percent correct	Listener's response						
			b	d	g	p	t	k	none
b	4	95.0	38			2			
d	17	88.2	6	150	2		12		
p	8	93.8	5			75			
t	15	88.7		17			133		
k	2	100.0						20	

Appendix C

Qualitative acoustic attributes

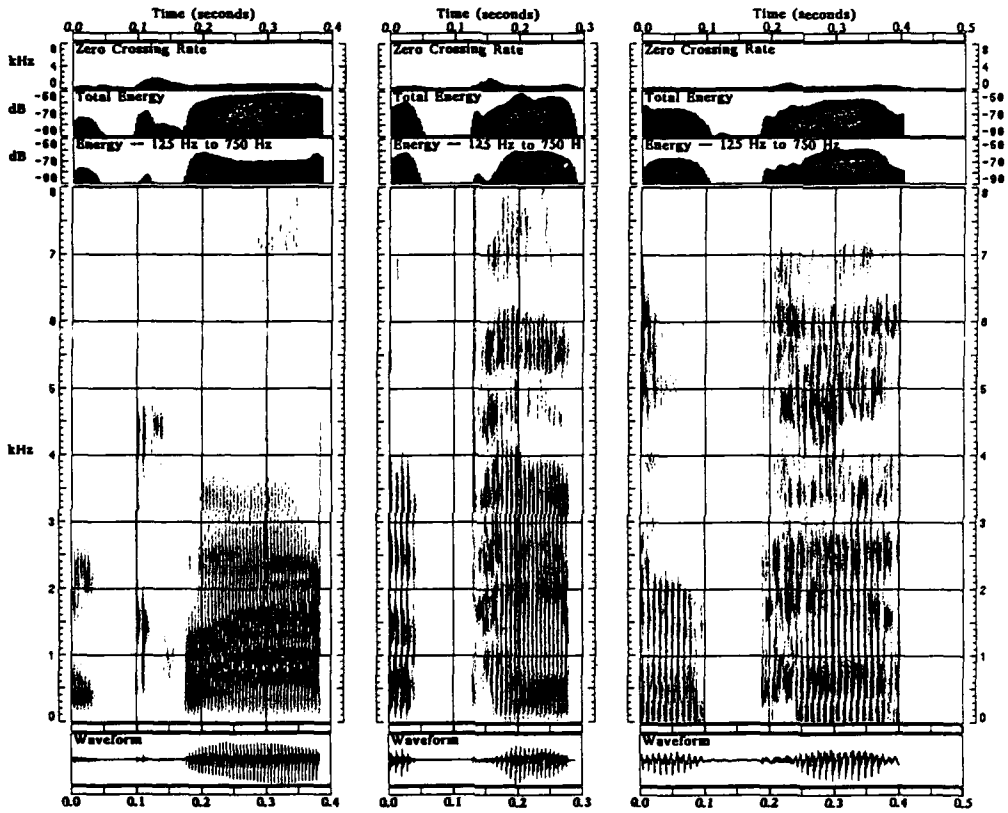


Qualitative acoustic attributes for voicing: (1) short VOT, maybe prevoiced, (2) mid VOT, prevoiced, F_1 motion into /ɛ/, (3) mid VOT, aspirated, and (4) long VOT, aspirated.

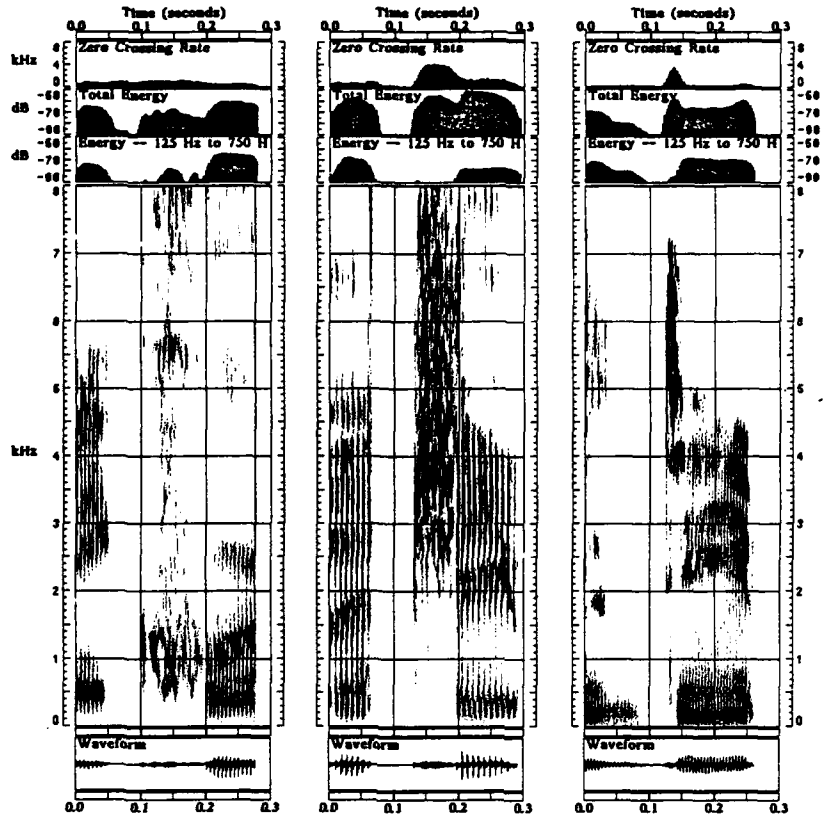


Qualitative acoustic attributes for the stop release frequency location: (5) low frequency [primarily below 2 kHz], (6) mid frequency [primarily in the range 2-4 kHz], (7) high frequency [primarily above 4 kHz], (8) bimodal [two concentrations of energy, the lower near F_2 and the higher at roughly three times the frequency of the lower], (9) broad [energy evenly distributed across all frequencies 0-8 kHz], and (10) no release visible in the spectrogram.

Appendix C. Qualitative acoustic attributes



(8) /ɨkwɑʸ/ (9) /ɘpeʸ/ 10) /ɜpæ/

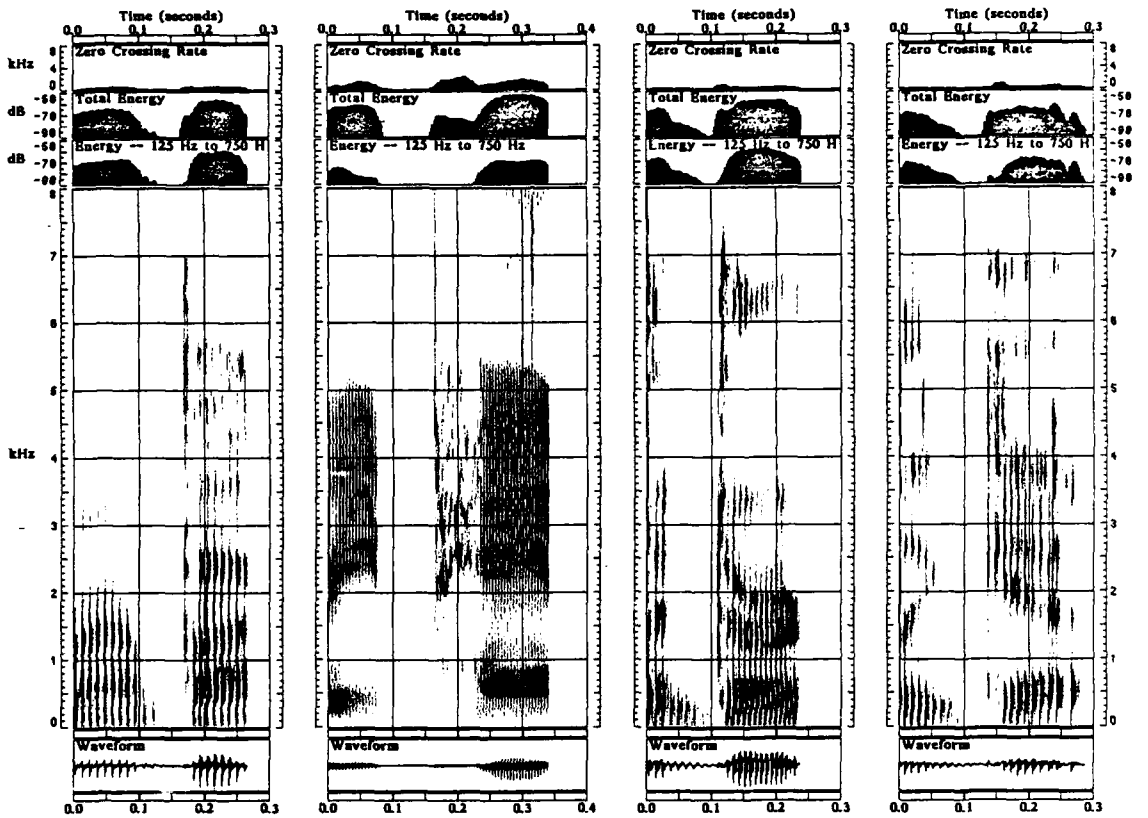


(11) /ikɔ/ (12) /əǰi/ (13) /ədi/

Qualitative acoustic attributes for the stop release relative to the formant frequency locations: (11) at and above F−4 with little energy below, (12) at and above F_3 with little energy below, (13) near F_2, with little energy above.

Qualitative acoustic attributes for the frequency distribution of energy in the release: even [even over all frequencies, 0-8 kHz] (14)(21), diffuse [over a frequency region of 2-4 kHz] (3)(7)(13), compact [in a frequency range of 1-1.5 kHz] (5)(11)(20), and bimodal [two concentrations of energy, the lower near F_2 and the higher at roughly three times the frequency of the lower] (4)(8)(20).

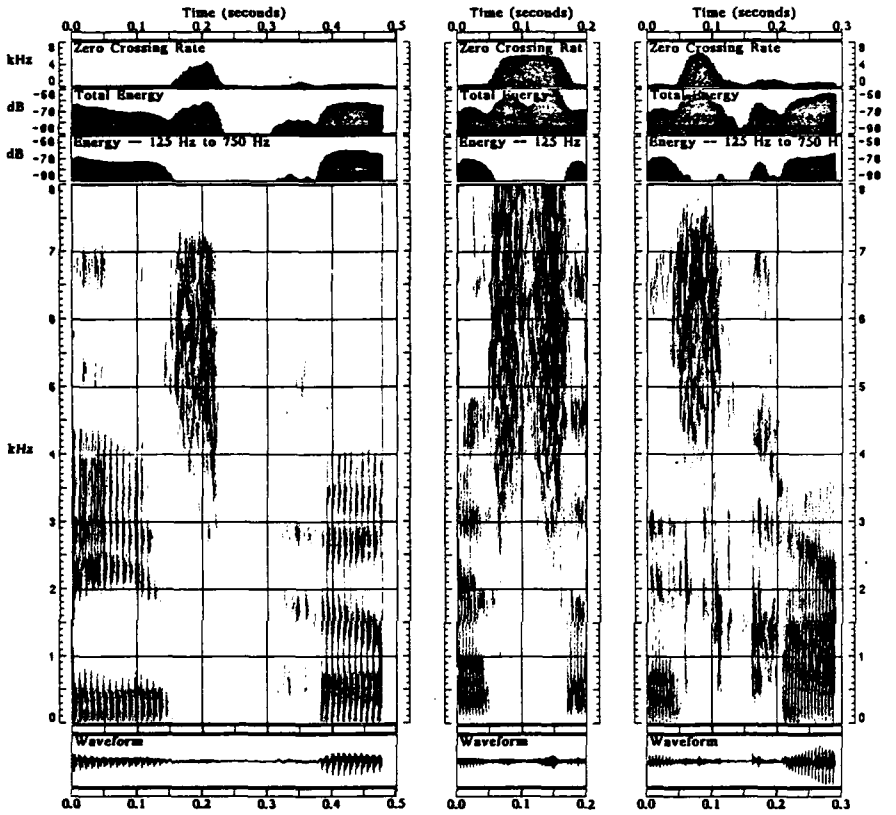
Qualitative acoustic attributes for strength of the release: weak [total energy is weak relative to the vowel] (1)(4)(10), and strong [total energy is strong relative to the vowel] (3)(7).



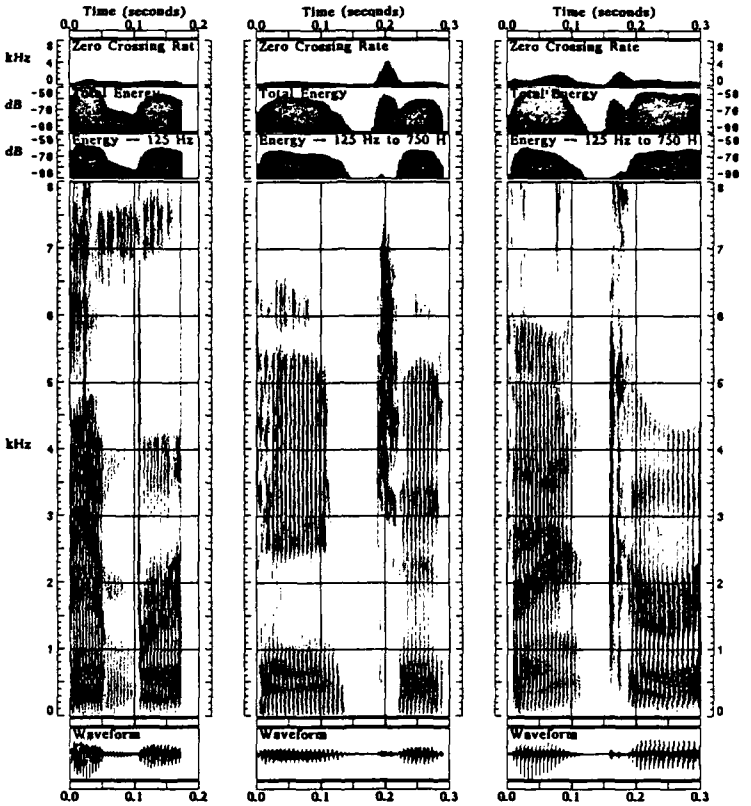
(14) /əbʌ/ (15) /ipe^{y}/ (16) /ɪdɝ/ (17) /igɪ/

Qualitative acoustic attributes for the motion of the second and third formants: (14) F_2 falling into the stop on the left and the right, (15) F_2 and F_3 falling into the stop on the left, and in the aspiration on the right,[even F_4 is falling on the left], (16) F_2 and F_3 rising into the stop from the right, the locus for F_2 is near 1800 Hz, (17) F_2 and F_3 come together into the stop, in what is referred to as a pinch.

Other, more subtle qualitative acoustic attributes: thin, pencil-like release (1)(2)(14), double-burst (4)(8)(14).



Qualitative acoustic attributes present in the fricative: (18) "labial-tail," lowering of the low-frequency energy limit of the fricative in anticipation of the stop, (19) incomplete closure between the fricative and the stop, (20) "velar blob," concentration of energy at the end of the fricative at roughly the same frequency location as the stop release.



(21) /ɛmbɔ/ (22) /eʸntɪ/ (23) /ɪŋkɔ/

Qualitative acoustic attributes present in the nasal: (21) long nasal murmur preceding voiced stop and (22,23) short nasal murmur preceding voiceless stops. Formant transitions into the vowel indicate the place of articulation of the nasal.

Appendix D

Rules

This appendix contains the rules used to identify stops. First an example of a backward-chaining query rule is given, the remainder of the query rules all have the same form and are not included. The next set of rules identify the voicing dimension, followed by rules to identify the place of articulation. The place rules have been divided into four parts: burst rules, formant transition rules, aspiration rules, and fricative rules. The final set of rules map from the vowel identity to vowel features. Weaker versions of most of the rules exist to handle uncertain evidence. Since these rules assert the same conclusions but with a weaker strength they have not been included.

```
;;; example of a query fule

(defrule query-prevoicing ""
  (declare (salience ?*voicing-query-salience2*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?token token)
    (instance-of ?token stop)
    (left-of ?left ?token)
    (class ?left vowel|nasal|semivowel)
    (goal (qualitative-attribute-of q-PREVOICED-yes ?closure)))
  (not (explicit (qualitative-attribute-of q-PREVOICED-yes|q-PREVOICED-no|q-PREVOICED-maybe ?closure)))
  (not (confirmed voicing-characteristic))
  => (bind ?answer #1(query-user ?closure 'prevoiced "Is there prevoicing during the closure interval? "))
  (and ?answer (assert (qualitative-attribute-of ?answer ?closure))))
```

```
;; VOT rules for voicing decision

;;; this rule probably shouldn't be certain, but as long as it needs to be confirmed, it's ok
; if the next vowel is reduced, then VOT-short may indicate voiceless, but weakly
(defrule voicing-VOT-short ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (split ((not (feature-of f-reduced ?right))
    => ; VOT short
    (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *with-certainty*)))
   ((feature-of f-reduced ?right)
    => ; VOT short, reduced
    (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *strong-evidence*)
    (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *weak-evidence*))))))

(defrule voicing-VOT-long ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-VOT-LONG-yes ?release))
  (not (confirmed voicing-characteristic))
  => ; VOT long
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *with-certainty*)))

; stops in semivowel clusters have longer VOT's
; therefore, if the VOT is medium and the stop is preceding a semivowel, it is likely to be voiced
(defrule voicing-VOT-cluster-not-short-or-long ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (right-of ?right ?token)
    (class ?right semivowel)
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release))
  (not (confirmed voicing-characteristic))
  => ; VOT not SHORT or LONG, in semivowel cluster
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *strong-evidence*)
  (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *weak-evidence*)))



;;; aspiration rules - some of these apply for all syllable-positions
;;;    aspirated-yes --> voiceless, independent of syllable-position
;;;       the inverse is not necessarily true

(defrule voicing-aspiration-yes-voiceless ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
```

```
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and ASPIRATED-yes
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *strong-evidence*)))

(defrule voicing-aspiration-no-voiced ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-ASPIRATED-no ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and ASPIRATED-no
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *medium-evidence*)))

; maybe aspiration and mid VOT may be a weak indicator for voiceless
(defrule voicing-aspiration-maybe-voiceless ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-ASPIRATED-maybe ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and ASPIRATED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence-maybe*)))

; aspiration and short may be a weak indicator for voiceless even if syllable-initial and unstressed
(defrule voicing-aspiration-reduced ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-VOT-SHORT-yes)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration)
    (right-of ?right ?token)
    (exists (feature-of f-reduced | f-syllabic ?right))) ; only fire once if both
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-yes and ASPIRATED-yes, reduced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))


; confirming rules for aspiration
(defrule confirm-voicing-aspiration-certain ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
```

```
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (exists (confirm voicing-characteristic ?token f-voiceless ?cf))
    (qualitative-attribute-of q-VOT-LONG-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; confirm: VOT LONG and ASPIRATED
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *with-certainty*)))

(defrule confirm-voicing-aspiration-medium ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (exists (confirm voicing-characteristic ?token f-voiceless ?cf))
    (qualitative-attribute-of q-VOT-LONG-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-maybe ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; confirm: VOT LONG and ASPIRATED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))

(defrule confirm-voicing-aspiration-no-certain ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (exists (confirm voicing-characteristic ?token f-voiced ?cf))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-no ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; confirm: VOT SHORT and not ASPIRATED
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *with-certainty*)))

(defrule ruleout-voicing-maybe-aspirated ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (exists (ruleout voicing-characteristic ?token f-voiced ?cf))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-maybe ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; ruleout: VOT SHORT and ASPIRATED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-negative-evidence*)))


;;;  for medium VOT's also need to look at more acoustic characteristics than just VOT
;;;  might want to consider these also for non-syllable-initial singleton stops

;;; rules for prevoicing
(defrule voicing-prevoicing-voiced-strong ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?closure closure)
```

```
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel) ; not applicable if preceded by stop or fricative
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-PREVOICED-yes ?closure))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and PREVOICED-yes
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *strong-evidence*)))

(defrule voicing-prevoicing-voiceless-medium ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-PREVOICED-no ?closure))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and PREVOICED-no
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))

(defrule voicing-prevoicing-voicing-weak-initial ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-PREVOICED-maybe ?closure))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and PREVOICED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence-maybe*)
  (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *strong-evidence-maybe*)))

; non-initial, intervocalic are more likely to be prevoiced if voiced
(defrule voicing-prevoicing-voicing-weak-non-initial ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release)
    (qualitative-attribute-of q-PREVOICED-maybe ?closure))
  (not (confirmed voicing-characteristic))
  => ; VOT SHORT-no, LONG-no and PREVOICED-maybe, non-initial
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)
  (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-evidence*)))
```

```
;;; confirming rules for prevoicing
;;; confirm that voicing is voiced
(defrule confirm-voicing-prevoicing-voiced-certain ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (exists (confirm voicing-characteristic ?token f-voiced ?cf))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-PREVOICED-yes ?closure))
  (not (confirmed voicing-characteristic))
  => ; confirm: VOT-SHORT-yes and PREVOICED-yes
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *with-certainty*)))

(defrule confirm-voicing-prevoicing-voiced-medium ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (exists (confirm voicing-characteristic ?token f-voiced ?cf))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-PREVOICED-maybe ?closure))
  (not (confirmed voicing-characteristic))
  (split (=> ; confirm: VOT-SHORT-yes and PREVOICED-maybe
     (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *medium-evidence*)))
 ((property-of syllable-non-initial|syllable-position-unknown ?token)
  => ; confirm: VOT-SHORT-yes and PREVOICED-maybe, non-initial
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *weak-evidence*)))))


;;; confirm voicing is voiceless
(defrule confirm-voicing-prevoicing-voiceless-certain-initial ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel | nasal | semivowel)
    (property-of syllable-initial ?token)
    (exists (confirm voicing-characteristic ?token f-voiceless ?cf))
    (qualitative-attribute-of q-VOT-LONG-yes ?release)
    (qualitative-attribute-of q-PREVOICED-no ?closure))
  (not (confirmed voicing-characteristic))
  => ; confirm: VOT-LONG-yes and PREVOICED-no
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *with-certainty*)))

(defrule ruleout-voicing-prevoicing-voiced-weak ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
```

```
    (class ?left vowel | nasal | semivowel)
    (not (property-of syllable-initial ?token))
    (exists (ruleout voicing-characteristic ?token f-voiced ?cf))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-PREVOICED-no ?closure))
  (not (confirmed voicing-characteristic))
  => ; ruleout: VOT short and not prevoiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-negative-evidence*)))


;;; voicing VOT rules for non-initial

(defrule voicing-non-initial-VOT-long ""
  (declare (salience ?*voicing-desperate*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (class ?right vowel)
    (not (property-of syllable-initial ?token))
    (qualitative-attribute-of q-VOT-LONG-yes ?release))
  (not (confirmed voicing-characteristic))
  => ; VOT long, non-initial
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence-maybe*)))

(defrule voicing-non-initial-VOT-not-short-or-long ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (class ?right vowel)
    (not (property-of syllable-initial ?token))
    (qualitative-attribute-of q-VOT-SHORT-no ?release)
    (qualitative-attribute-of q-VOT-LONG-no ?release))
  (not (confirmed voicing-characteristic))
  => VOT not SHORT or LONG, non-initial
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-evidence-maybe*)
  (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence-maybe*)))


;;; aspiration rules for non-initial
; these are independent of VOT
; the presence of aspiration indicates voiceless, but
; the lack of aspiration does not necessarily indicate voiced!

(defrule voicing-non-initial-aspiration-yes ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (not (property-of syllable-initial ?token))
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; non-initial: ASPIRATED-yes
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *strong-evidence*)))

(defrule voicing-non-initial-aspiration-maybe ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
```

```
    (instance-of ?release release)
    (interval-of ?release ?token)
    (not (property-of syllable-initial ?token))
    (qualitative-attribute-of q-ASPIRATED-maybe ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; non-initial: ASPIRATED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))


; these aspiration rules are secondary, and only fire after VOT is known
; if the stop is not syllable-initial, then any aspiration indicates voiceless? (even if VOT is short)
(defrule voicing-non-initial-short-aspiration-yes ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (not (property-of syllable-initial ?token))
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; non-initial: VOT SHORT-yes and ASPIRATED-yes
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))

(defrule voicing-non-initial-short-aspiration-maybe ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-non-initial|syllable-position-unknown ?token)
    (qualitative-attribute-of q-VOT-SHORT-yes ?release)
    (qualitative-attribute-of q-ASPIRATED-maybe ?aspiration))
  (not (confirmed voicing-characteristic))
  => ; non-initial: VOT SHORT-yes and ASPIRATED-maybe
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence-maybe*)))


;;; voicing from f1 motion

(defrule voicing-F1-motion-right ""
  (declare (salience ?*voicing-desperate*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F1-FALLING-yes ?right)
    (qualitative-attribute-of q-F1-AMT-LARGE-yes ?right))
  (not (confirm voicing-characteristic ?token ?vcg ?cf))
  (not (confirmed voicing-characteristic))
  => ; lots of RF1 motion favors voiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-evidence*)))

(defrule voicing-F1-motion-left ""
  (declare (salience ?*voicing-desperate*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left vowel|semivowel)
```

```
    (qualitative-attribute-of q-DURATION-SHORT-no ?left)
    (qualitative-attribute-of q-F1-FALLING-yes ?left)
    (qualitative-attribute-of q-F1-AMT-LARGE-yes ?left))
  (not (confirm voicing-characteristic ?token ?vcg ?cf))
  (not (confirmed voicing-characteristic))
  => ; lots of LF1 motion favors voiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *weak-evidence*)))


;;; voicing for non-initial from total stop duration

(defrule voicing-non-initial-duration-short ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (qualitative-attribute-of q-DURATION-SHORT-yes ?token))
  (not (property-of syllable-initial ?token))
  (not (confirmed voicing-characteristic))
  => ; duration short --> voiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *medium-evidence*)))


(defrule voicing-from-semivowel-duration-long ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (right-of ?right ?token)
    (class ?right semivowel)
    (qualitative-attribute-of q-DURATION-LONG-yes ?right))
  (not (confirmed voicing-characteristic))
  => ; LONG semivowel --> voiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiced *medium-evidence*)))

(defrule voicing-from-semivowel-duration-short ""
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (not (property-of s-cluster ?token))
    (right-of ?right ?token)
    (class ?right semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-yes ?right))
  (not (confirmed voicing-characteristic))
  => ; LONG semivowel --> voiced
  (assert (add-to-score =(inccntr) ?token voicing-characteristic f-voiceless *medium-evidence*)))
```

```
;;; these rules all use features of the vowel derived from the vowel id
;;; another version uses formant locations directly and a third tries
;;; to derive the vowel features

;;;  rules for place identification

; primary burst rules

(defrule place-burst-not-visible ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-BURST-VISIBLE-no ?release))
  (not (confirmed place-of-articulation))
  => ; place: burst not-visible
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)))

(defrule place-burst-broad ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (qualitative-attribute-of q-BURST-LOCATION-BROAD-yes ?release))
  (not (confirmed place-of-articulation))
  (split ((property-of syllable-initial ?token)
          => ; place: burst-broad
          (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)))
         ((not (property-of syllable-initial ?token))
          => ; place: burst-broad
          (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)
          (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))))

(defrule place-pencil-like ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
  ; check that there is already some belief in labial
  ; ie. don't ask unless there is some reason to believe it!
    (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-labial ?amt&:(> (eval ?amt) 0)))
    (qualitative-attribute-of q-THIN-yes ?release))
  (not (confirmed place-of-articulation))
   => ; place: pencil-like
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)))

(defrule place-HF ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-LOCATION-HF-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((class ?right vowel)
         (feature-of f-front ?right)
         (not (feature-of f-round|f-retroflex ?right))
        => ; place: HF, front vowel
        (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)
        (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*))))
  => ; place: HF
```

```
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence*)))

; ask this only if deciding between labial and alveolar
(defrule place-wedge-like ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
  ; check that there is already some belief in labial and some in alveolar
  ; ie. don't ask unless there is some reason to believe it!
    (exists (add-to-score ?uniqueness-number1 ?token place-of-articulation f-labial ?amt1&:(> (eval ?amt1) 0)))
    (exists (add-to-score ?uniqueness-number2 ?token place-of-articulation f-alveolar ?amt2&:(> (eval ?amt2) 0)))
    (qualitative-attribute-of q-THICK-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: wedge-like
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)))

(defrule place-MF ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-BURST-LOCATION-MF-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-retroflex ?right)
        => ; place: MF, retroflex vowel
        (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence*)
        (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
((feature-of f-round ?right)
 => ; place: MF, round vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence*)))
((feature-of f-back ?right)
 => ; place: MF, back vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)))
((feature-of f-front ?right)
 => ; place: MF, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))
((feature-of f-schwa ?right)) ; schwa could be underlying round or retro
=> place: MF, schwa ; this should be a weak assertion
(assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*))))

(defrule place-LF ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (right-of ?right ?token)
    (qualitative-attribute-of q-BURST-LOCATION-LF-yes ?release)
    (class ?right vowel|semivowel)
    (not (confirmed place-of-articulation))
    (case ((exists (feature-of f-round|f-retroflex ?right))
  => ; place: low, round|retroflex vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)
  (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)
  (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
  ((feature-of f-back ?right)
```

```
    => ; place: low, back vowel
    (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)
    (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
   ((feature-of f-front ?right)
    => place: low, front vowel
    (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-negative-evidence*)
    (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-negative-evidence*))))))

(defrule place-location-bimodal ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial ?token)
    (right-of ?right ?token)
    (class ?right vowel | semivowel)
    (not (exists (feature-of f-front ?right)))
    (qualitative-attribute-of q-BURST-LOCATION-BIMODAL-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: bimodal, back vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))


;;; relative (to formant locations) location rules
(defrule place-below-F2 ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-BELOW-F2-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: burst below F2
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))

(defrule place-compact-at-F2 ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-BIMODAL-no ?release) ; inhibit this rule if bimodal
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-AT-F2-yes ?release)
    (not (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-COMPACT-no ?release))
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (feature-of f-back ?right))
  (not (confirmed place-of-articulation))
  => ; place: compact at F2, back vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)
  (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))

(defrule place-compact-above-F2 ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-BIMODAL-no ?release) ; inhibit this rule if bimodal
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-ABOVE-F2-yes ?release)
    (not (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-COMPACT-no ?release)))
  (not (confirmed place-of-articulation))
  => ; place: burst above F2
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))
```

```
(defrule place-above-F3 ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-ABOVE-F3-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: burst above F3
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))

(defrule place-above-F4 ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-ABOVE-F4-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: burst above F4
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence*)))

(defrule place-relative-location-bimodal ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (right-of ?right ?token)
    (class ?right vowel | semivowel)
    (not (exists (feature-of f-front ?right)))
    (qualitative-attribute-of q-BURST-RELATIVE-LOCATION-BIMODAL-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: burst bimodal, back vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))

(defrule place-double-burst ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial ?token)
    ; check that there is already some belief in velar
    (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-velar ?amt&:(> (eval ?amt) 0)))
    (qualitative-attribute-of q-DOUBLE-BURST-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: double burst
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))

;;; specific rule to differentiate fron t and k. the front k's sometimes have
;;; a weakening in the release, giving almost a bimodal appearance, due to a zero
(defrule place-top-spectral-zero ""
  (declare (salience ?*confirm-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (feature-of f-front ?right))
  (not (confirmed place-of-articulation))
  (qualitative-attribute-of q-BURST-LOCATION-HF-yes|q-BURST-LOCATION-HF-maybe ?release)
```

```
  ;  only  if can't decide between front velar and front alveolar
  (top-candidate place-of-articulation ?token f-velar|alveolar ?amt1&:(> ?amt1 0))
  (2nd-candidate place-of-articulation ?token f-velar|alveolar ?amt2&:(> ?amt2 0))
  (split ((qualitative-attribute-of q-SPECTRAL-ZERO-yes ?release)
  => ; place top: spectral zero in release
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
 ((qualitative-attribute-of q-SPECTRAL-ZERO-no ?release)
  => ; place top: no spectral zero in release
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))))


;;; energy distribution rules
;;; compact/diffuse rules
(defrule place-energy-compact ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-COMPACT-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((class ?right vowel|semivowel)
 (exists (feature-of f-round|f-retroflex|f-lateral ?right)))
=> place: compact, round or retroflex
(assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)
(add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))
  => ; place: compact
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))

(defrule place-energy-diffuse-HF-or-MF ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (exists (qualitative-attribute-of q-BURST-LOCATION-HF-yes|q-BURST-LOCATION-HF-maybe|
      q-BURST-LOCATION-MF-yes|q-BURST-LOCATION-MF-maybe ?release))
    (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-DIFFUSE-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right))
=> ; place: diffuse Hf or MF, front vowel
(assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
  => ; place: diffuse HF or MF
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)
  (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)))

(defrule place-energy-diffuse-LF ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (right-of ?right ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-LOCATION-LF-yes|q-BURST-LOCATION-LF-maybe ?release)
    (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-DIFFUSE-yes ?release))
  (not (confirmed place-of-articulation))
  =>   place: diffuse LF
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))

; for back vowels the bimodal nature is 1/4 3/4
; for front vowels it is more 1/2 wavelength.
(defrule place-energy-bimodal ""
```

```
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (property-of syllable-initial | syllable-non-initial | s-cluster ?token)
    (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-BIMODAL-yes ?release))
  (not (confirmed place-of-articulation))
  => ; place: energy distribution bimodal
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))

(defrule place-energy-even-initial ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (qualitative-attribute-of q-BURST-ENERGY-DISTRIBUTION-EVEN-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((property-of syllable-initial ?token)
 => ; place: energy distribution even
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)))
((not (property-of syllable-initial ?token))
 => ; place: energy distribution even
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))))


;;; strength rules

(defrule place-strength-strong-initial ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (qualitative-attribute-of q-STRENGTH-STRONG-yes ?release)
    (not (property-of semivowel-cluster ?token)))        ; hard to assess strength in clusters
  (not (confirmed place-of-articulation))
  (case ((property-of syllable-initial ?token)
 => ; place: release strong
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))
((property-of syllable-non-initial|s-cluster ?token)
 => ; place: release strong
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))))

(defrule place-strength-weak-initial ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (qualitative-attribute-of q-STRENGTH-WEAK-yes ?release))
  (not (confirmed place-of-articulation))
  (case ((property-of syllable-initial ?token)
 => ; place: release weak
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
((property-of syllable-non-initial|s-cluster ?token)
 => place: release weak
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)))))

; aspiration strength
(defrule place-aspiration-strength "strong aspiration and weak release supports labial"
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
```

```
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (property-of syllable-initial ?token)
    (qualitative-attribute-of q-STRENGTH-WEAK-yes ?release)
    (qualitative-attribute-of q-STRENGTH-STRONG-yes ?aspiration))
  (not (confirmed place-of-articulation))
  => ; place: release weak, aspiration strong
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
```

```
;;;  rules are shown for the right-context, the same rules exist for the left

;;;  formant rules -- rules based on features, specified by identity
;;; if the vowel is short, then don't use formant motion because it may
;;; be misleading, only use combined motion rules as low priority

(defrule formants-F2-rising-large ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-RISING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-LARGE-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((exists (feature-of f-back|f-round ?right))
 =>  place: RF2 rising into stop, back or round vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence*)))
    ((feature-of f-front ?right)
     => place: RF2 rising into stop, front vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-negative-evidence*)))
    (otherwise
      => ; place: RF2 rising into stop, vowel unknown
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)))))

(defrule formants-F2-rising-large-fb ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (left-of ?left ?token)
    (class ?left vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-RISING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-LARGE-yes ?right))
  (not (confirmed place-of-articulation))
  (exists (feature-of f-back|f-round ?right))
  (case ((feature-of f-front ?left)
 => ; place: RF2 rising into stop, back or round vowel, left front
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
    (otherwise
 => ; place: RF2 rising into stop, back or round vowel, left front
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)))))

(defrule formants-RF2-rising-medium ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-RISING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-LARGE-maybe ?right))
  (not (confirmed place-of-articulation))
  (case ((exists (feature-of f-back|f-round ?right))
 => ; place: RF2 rising into stop, back or round vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence-maybe*)
```

```
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *strong-evidence-maybe*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence-maybe*)))
    ((feature-of f-front ?right)
     => ; place: RF2 rising into stop, front vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence-maybe*)
     (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-negative-evidence-maybe*)))
    (otherwise
      => ; place: RF2 rising into stop, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence-maybe*)))))

(defrule formants-RF2-rising-small ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-RISING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-SMALL-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((exists (feature-of f-back|f-round ?right))
 => ; place: RF2 rising into stop, back or round vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-negative-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)))
    ((feature-of f-front ?right)
     => ; place: RF2 rising into stop, front vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-negative-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))
    (otherwise
      => ; place: RF2 rising into stop, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-negative-evidence*)))))


; F2 falling
(defrule formants-RF2-falling-large ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-FALLING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-LARGE-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 => ; place: RF2 falling into stop, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-negative-evidence*)))
    ((feature-of f-back ?right)
     => ; place: RF2 falling into stop, back vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-negative-evidence*)))
    (otherwise
      => ; place: RF2 falling into stop, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))))

(defrule formants-RF2-falling-medium ""
  (declare (salience ?*formant-salience*))
  (logical
```

```
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-FALLING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-LARGE-maybe ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 => ; place: RF2 falling into stop, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence-maybe*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence-maybe*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-negative-evidence-maybe*)))
    ((feature-of f-back ?right)
     => ; place: RF2 falling into stop, back vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence-maybe*)
     (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence-maybe*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-negative-evidence-maybe*)))
    (otherwise
      => ; place: RF2 falling into stop, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence-maybe*)))))

(defrule formants-RF2-falling-small ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-FALLING-yes ?right)
    (qualitative-attribute-of q-F2-AMT-SMALL-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 => ; place: RF2 falling into stop, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-negative-evidence*)))
    ((feature-of f-back ?right)
     => ; place: RF2 falling into stop, back vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *strong-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-negative-evidence*)))
    (otherwise
      => ; place: RF2 falling into stop, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)))))

(defrule formants-RF2-flat ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-RISING-no ?right)
    (qualitative-attribute-of q-F2-FALLING-no ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 (feature-of f-high ?right)
 (not (feature-of f-reduced ?right))
 => ; place: RF2 flat, high,front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-negative-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)))
```

```
    ((feature-of f-front ?right)
     (not (feature-of f-reduced ?right))
     => ; place: RF2 flat, front vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-negative-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))
    ((feature-of f-back ?right)
     (not (feature-of f-reduced ?right))
     => ; place: RF2 flat, back vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-negative-evidence*)))
    (otherwise
      => ; place: RF2 flat, vowel unknown
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)))))

; F3 rising
(defrule formants-RF3-rising ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F3-RISING-yes ?right))
  (not (confirmed place-of-articulation))
  (split ( => ; place: RF3 rising into stop
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)))
 ((exists (feature-of f-round|f-retroflex ?right))
  => ; place: RF3 rising into stop, vowel round or retroflex
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))
 ((feature-of f-front ?right)
  => ; place: RF3 rising into stop, front vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))))

; F3 falling
(defrule formants-RF3-falling ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F3-FALLING-yes ?right))
  (not (confirmed place-of-articulation))
  (split ( =>  ; place: RF3 falling into stop
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
 ((feature-of f-retroflex ?right)
  => ; place: RF3 falling into stop, retro vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-negative-evidence*)))))


;;; combined formant motion rules

(defrule formants-right-pinch ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-F3-PINCH-yes ?right))
```

```
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 (not (feature-of f-round ?right))
 => ; right F2 and F3 pinch
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *strong-evidence*)))
    ((feature-of f-round ?right)
     => ; right F2 and F3 pinch, round
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*))))
  => (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-negative-evidence*)))

(defrule formants-no-right-pinch ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-F3-PINCH-no ?right)
    (feature-of f-front ?right)
    (not (feature-of f-round ?right)))
  (not (confirmed place-of-articulation))
    => ; right F2 and F3 do not pinch
    (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *weak-negative-evidence*)))

(defrule formants-falling-right ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-FORMANTS-ALL-FALLING-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 (feature-of f-high ?right)
 => ; formants all falling on the right, high, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)
 (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
    ((feature-of f-front ?right)
     => ; formants all falling on the right, front vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *weak-evidence*)
     (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
    (otherwise
      => ; formants all falling on the right
      (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))))

(defrule formants-rising-right ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-DURATION-SHORT-no ?right)
    (qualitative-attribute-of q-F2-F3-RISING-yes ?right))
  (not (confirmed place-of-articulation))
  (case ((feature-of f-front ?right)
 => ; F2 and F3 rising on the right, front vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
    ((feature-of f-back ?right)
     => ; F2 and F3 rising on the right, back vowel
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))
```

```
((feature-of f-round ?right)
 => ; F2 and F3 rising on the right, round vowel
 (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))))
```

```
;;;   rules for the aspiration characteristics

(defrule aspiration-F2-hole ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (feature-of f-front ?right)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration))
  (not (confirmed place-of-articulation))
  ;  only used to decide between front velar and front alveolar
  (exists (add-to-score ?uniqueness-number1 ?token place-of-articulation f-velar ?amt1&:(> (eval ?amt1) 0)))
  (exists (add-to-score ?uniqueness-number2 ?token place-of-articulation f-alveolar ?amt2&:(> (eval ?amt2) 0)))
  (split ((qualitative-attribute-of q-F2-HOLE-yes ?aspiration)
  => ; F2 hole in aspiration
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))
 ((qualitative-attribute-of q-AT-F2-yes ?aspiration)
  => ; aspiration at F2
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))))

(defrule aspiration-labial-tail ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration)
    (qualitative-attribute-of q-TAIL-yes ?aspiration))
  (not (confirmed place-of-articulation))
  (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-labial ?amt&:(> (eval ?amt) 0)))
  => ; labial tail in aspiration
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))

(defrule aspiration-AT-F2-only ""
  (declare (salience ?*formant-salience*))
  (logical
    (instance-of ?token token)
    (class ?token stop)
    (instance-of ?aspiration aspiration)
    (interval-of ?aspiration ?token)
    (right-of ?right ?token)
    (class ?right vowel|semivowel)
    (qualitative-attribute-of q-ASPIRATED-yes ?aspiration)
    (qualitative-attribute-of q-AT-F2-ONLY-yes ?aspiration))
  (not (confirmed place-of-articulation))
  (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-labial ?amt&:(> (eval ?amt) 0)))
  (split (=> ; aspiration at F2 only
     (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))
 ((feature-of f-back ?right)
  => ; aspiration at F2 only, back vowel
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))))
```

```
; after the user specifies the identity of the fricative,
; make it an instance of that fricative
(defrule fricative-from-identity ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token fricative)
    (asked-identity ?token ?id)
    (value-of ?fricative fricative)
    (explicit (identity ?fricative ?id)))
  => (assert (instance-of ?token ?fricative)))

; give an instance of a fricative all the features of its fricative class
(defrule fricative-features ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token fricative)
    (instance-of ?token ?fricative)
    (value-of ?fricative fricative)
    (explicit (feature-of ?feature ?fricative)))
  => (assert (feature-of ?feature ?token)))

;;; rules to try and identify the fricative as an s or a z
(defrule fricative-striations ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token stop)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left fricative)
    (qualitative-attribute-of q-STRIATIONS-yes ?left))
  => ; fricative has striations, so is likely to be a z
  (assert (property-of s-not-cluster ?token)))

(defrule fricative-vbar ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token stop)
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left fricative)
    (qualitative-attribute-of q-VBAR-yes ?left))
  => ; fricative has a VBAR, so is likely to be a z
  (assert (property-of s-not-cluster ?token)``


;;; if preceded by fricative, and VOT is mid and incomplete closure, then
;;; the voicing of the stop and the fricative are probably the same
(defrule fricative-stop-voicing-agree "rules to deduce the voicing characteristic"
  (declare (salience ?*voicing-rule-salience*))
  (logical
    (instance-of ?token stop)
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (left-of ?left ?token)
    (class ?left fricative)
    (voicing-characteristic ?left ?vc)
    (qualitative-attribute-of q-INCOMPLETE-CLOSURE-yes ?closure))
  (not (confirmed voicing-characteristic))
  => ; incomplete closure --> voicing agrees
  (assert (add-to-score =(inccntr) ?token voicing-characteristic ?vc *medium-evidence*)))
```

```
;;; the context rules for formants are no longer applicable,
;;; and are inhibited because the left context of the stop is no longer
;;; a vowel

(defrule s-burst-not-visible ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?release release)
    (interval-of ?release ?token)
    (left-of ?left ?token)
    (class ?left fricative)
    (property-of s-cluster ?token)
    (qualitative-attribute-of q-BURST-VISIBLE-no ?release))
  (not (confirmed place-of-articulation))
  => ; fricative place: burst not-visible
  (assert (add-to-score =(inccntr) ?token place-of-articulation f-labial *weak-evidence*)))

;;;  place rules for fricative-clusters: cues in the fricative --- do not require cluster
(defrule fricative-spectral-tail ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token stop)
    (left-of ?left ?token)
    (class ?left fricative))
  (not (confirmed place-of-articulation))
  (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-labial ?amt&:(> (eval ?amt) 0)))
  (qualitative-attribute-of q-CUTOFF-FALLING-yes ?left)
  => ; fricative place: spectral-tail
  (assert  (add-to-score =(inccntr) ?token place-of-articulation f-labial *medium-evidence*)))

(defrule fricative-F2-blob ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token stop)
    (left-of ?left ?token)
    (class ?left fricative))
  (not (confirmed place-of-articulation))
  (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-velar ?amt&:(> (eval ?amt) 0)))
  (qualitative-attribute-of q-F2-BLOB-yes ?left)
  => ; fricative place: F2-blob
  (assert  (add-to-score =(inccntr) ?token place-of-articulation f-velar *medium-evidence*)))

(defrule fricative-incomplete-closure ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token stop)
    (instance-of ?closure closure)
    (interval-of ?closure ?token)
    (left-of ?left ?token)
    (class ?left fricative))
  (not (confirmed place-of-articulation))
  (exists (add-to-score ?uniqueness-number ?token place-of-articulation f-alveolar ?amt&:(> (eval ?amt) 0)))
  (qualitative-attribute-of q-INCOMPLETE-CLOSURE-yes ?closure)
  => ; fricative place: incomplete-closure
  (assert  (add-to-score =(inccntr) ?token place-of-articulation f-alveolar *medium-evidence*)))
```

```
;;; vowel rules


; after the user specifies the identity of the vowel, make it an instance
;  of that vowel
(defrule vowel-from-identity ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token vowel)
    (asked-identity ?token ?id)
    (value-of ?vowel vowel)
    (explicit (identity ?vowel ?id)))
  => (assert (instance-of ?token ?vowel)))

; give an instance of a vowel all the features of its vowel class
(defrule vowel-features ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token vowel)
    (instance-of ?token ?vowel)
    (value-of ?vowel vowel)
    (explicit (feature-of ?feature ?vowel)))
  => (assert (feature-of ?feature ?token)))

;;; features for the diphthongs depend on which side of stop
(defrule vowel-features-diphthong-left ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token vowel)
    (right-of ?right ?token)
    (instance-of ?right stop)
    (instance-of ?token vowel-ay|vowel-cy))
  => (assert (feature-of f-front ?token)))

(defrule vowel-features-diphthong-right ""
  (declare (salience ?*context-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token vowel)
    (left-of ?left ?token)
    (instance-of ?left stop)
    (split ((instance-of ?token vowel-ay)
   => (assert (feature-of f-back ?token)))
   ((instance-of ?token vowel-cy)
   => (assert (feature-of f-back ?token)
      (feature-of f-round ?token))))))

;;; semivowel features
(defrule semivowel-from-identity ""
  (declare (salience ?*place-rule-salience*))
  (logical
    (instance-of ?token token)
    (instance-of ?token semivowel)
    (asked-identity ?token ?id)
    (value-of ?semivowel semivowel)
    (explicit (identity ?semivowel ?id)))
  => (assert (instance-of ?token ?semivowel)))

(defrule semivowel-features ""
  (declare (salience ?*context-salience*))
  (logical
```

```
   (instance-of ?token token)
   (instance-of ?token semivowel)
   (instance-of ?token ?semivowel)
   (value-of ?semivowel semivowel)
   (explicit (feature-of ?feature ?semivowel)))
 => (assert (feature-of ?feature ?token)))
```

END